

A Thesis Submitted for the Degree of PhD at the University of Warwick

Permanent WRAP URL:

<http://wrap.warwick.ac.uk/88724>

Copyright and reuse:

This thesis is made available online and is protected by original copyright.

Please scroll down to view the document itself.

Please refer to the repository record for this item for information to help you to cite it.

Our policy information is available from the repository home page.

For more information, please contact the WRAP Team at: wrap@warwick.ac.uk

Intentional Agency

by

Zack Evans

Thesis submitted in partial fulfilment of the
requirements for the degree of

Doctor of Philosophy

University of Warwick
Department of Philosophy

August 2016

Table of Contents

List of Illustrations	5
Acknowledgements	6
Declaration	7
Abstract	8
1. Introduction	9
1.1 The Problem	9
1.2 Overview of the Argument	14
2. The Problem of ‘Free Will’	22
2.1 Introduction	22
2.2 Free Will as a Condition of Moral Responsibility	28
2.2.1 Incompatibilism	29
2.2.2 Compatibilism and Revisionism	33
2.3 Moving Beyond ‘Free Will’	40
2.3.1 Objections to my view	41
2.3.2 Fischer and Semicompatibilism	46
2.4 Conclusion	59
3. Control and Sourcehood	62
3.1 Introduction	62
3.2 Sourcehood and Agent Causation	65
3.3 Source Incompatibilism	75
3.3.1 The Four-Case Manipulation Argument	79
3.3.2 Responding to the Four-Case Argument	87

3.5 Agent Causation	102
3.6 Conclusion	116
4. Folk Psychology and Agency	120
4.1 Introduction	120
4.2 Agency Incompatibilism	127
4.2.1 The folk theory of mind and the concept of agency	129
4.2.2 ‘Settling’ and Agency Incompatibilism	135
4.2.3 Experimental evidence for an incompatibilist concept of agency	140
4.2.4 Intuitions about the concept of agency	148
4.3 The Evidentiary Role of Folk Psychology and Intuition	158
4.3.1 Debunking	159
4.3.2 The relation between cognitive science and metaphysics	169
4.4 Conclusion	179
5. Intentional Realism	181
5.1 Introduction	181
5.2 The Agency Concept	184
5.3 Challenges to Intentional Realism	198
5.3.1 Eliminativism	201
5.3.2 Instrumentalism	206
5.4 Interventionism and Mental Causation	215
5.4.1 Causation and Control Variables	224

5.5 Conclusion	239
6. Consciousness in Action	241
6.1 Introduction	241
6.2 Intervening on Beliefs	245
6.3 Conscious Perceptual Experience and ‘Zombie Action’	250
6.3.1 Perceptual demonstratives and the ‘two streams’ view	255
6.3.2 Embodied demonstratives	271
6.4 Disputing the Evidence for Zombie Action	277
7. Conclusion	289
7.1 The Problem Reconsidered	289
7.2 Causal Integration as Central to Agency	291
Bibliography	296

List of Illustrations

Figure 3.1: ‘Plum in Normal Deterministic World’	93
Figure 3.2: ‘Plum Subject to Manipulation’	93
Figure 6.1 a-b: ‘Ebbinghaus Ring Illusion’	260

Acknowledgements

It is impossible to identify, let alone acknowledge the individual contribution of, all of the casually relevant factors that influenced this thesis. Christoph Hoerl was my thesis advisor and one such causally relevant factor. No doubt there are numerous philosophical interlocutors, living or otherwise, whose influence can be seen in the present work.

Acknowledgement should be given to them — it may be more obvious to the reader than the author whom exactly they are. As the joke goes, any mistakes or particularly egregious errors in the following work should of course be attributed to them, and I trust that the reader can find enough information to track them down and hold them accountable.

Personal thanks should be given to my wife, for being supportive of me while I have been engaged in the quite peculiar labour of writing a Ph.D thesis in philosophy. And acknowledgement must also go to my daughter — for simply being. Personal and genetic thanks go to my parents for obvious reasons.

Finally, no ‘determinist-friendly’ work like this would be complete without acknowledging the contribution that is due to the causal summation of the entire history of the universe. (Just in case.)

Declaration

This thesis is my own work, and is not the result of collaborative research. No part of this work has been previously published. This thesis has not been submitted for any qualification at any other university.

Abstract

There are two central arguments in this project. The first is a kind of ‘second-order’ argument, that is, an argument *about* the dialectical situation of an existing argument (namely, about the ‘free will problem’). The second is a straightforward argument about agency, but one which can be better addressed—I claim—once the second-order argument has been made.

The ‘free will problem’ is widely claimed to be one of the perennial philosophical problems. But it is not one that has any widely accepted solution. The reason for this, as others have acknowledged, is due in large part to the wide range of problems that have historically been considered under the rubric of ‘the free will problem’.

My proposal is straightforward: stop talking about ‘free will’ altogether! More precisely, my claim is that we could—in principle—eliminate the term. However, it may be more difficult *in practise* to actually cease using the term, and so my prescription is to define the term operationally, as a philosophers’ technical term. As I will go on to explain, ‘free will’ means something like: ‘whatever it is, if anything, in virtue of which people are appropriate subjects of moral responsibility’.

The second argument then becomes apparent: setting aside the question of moral responsibility, we can see that there are a number of putative ‘free will’ issues that don’t go away. While most things can be sectioned off into the moral responsibility debate, as explained above, several of these issues actually turn out to depend on the concept of *agency*. This has not previously been recognised because of the structure of the ‘free will debate’, and especially because of its fixation on the notions of determinism and indeterminism. I then go on to sketch the outlines of a positive account of agency that can independently address those concerns which were previously thought to be about ‘free will’.

1

Introduction

1.1 The Problem

There are two central arguments in this project. The first of these can best be thought of as a ‘second-order’ argument: it is an argument *about* the dialectical state of an existing philosophical argument, or group of arguments — namely, the ‘free will problem’. The second is a straightforward argument about agency. However, part of my claim is that the theory of agency that I am proposing is best addressed once the second-order argument has been made. I will address these in turn.

The ‘free will problem’ is widely claimed to be one of the perennial philosophical problems. But it is not one that has any widely accepted solution. The reason for this, as others have acknowledged, is due in large part to the wide range of philosophical issues that have historically been addressed under the rubric of ‘the free will problem’. These issues could involve

anything from worries about divine foreknowledge to the philosophical implications of the latest work in neuroscience and neurobiology on ‘free will’.¹

In the following chapters, I begin by addressing the current state of the ‘free will debate’. One of the principal arguments I make is that, strictly speaking, it would be possible to eliminate the term ‘free will’ from the philosophical vocabulary, without significant loss of substance. Nearly all of the important, philosophically substantial discussions could carry on without the term: anything that could not be thus continued is likely to be confused, perhaps only a ‘verbal dispute’.²

With that said, I will not make much of the ‘verbal dispute’ claim. It is in any case probable that wholesale elimination of the term ‘free will’ is more trouble than it is worth, for reasons of clarity and convenience. The point of emphasising the theoretical possibility of its elimination is to support my claims about what ‘the free will problem’ is *really* about. I suggest that, in most cases,

¹ See John Fischer’s (1989) edited volume as an example of the former. As for the latter, there has been philosophical interest—one way or the other—in what brain science means for ‘human freedom’ almost as long as there has been a science of the brain. For a recent example, see Tse (2013).

² For some discussion of the notion of a ‘verbal dispute’ see Chalmers (2011).

what is really at issue is the nature of moral responsibility — and it is quite possible to recognise this and formulate all of the philosophically important research questions in such terms (i.e. questions about the various conditions or properties that one might think are required for this or that theory of responsibility).

Now, there are two points to clarify here. First, I said it is *mostly* the case that moral responsibility is at issue. There are certain areas where this is not true, and these cases are in fact more interesting to me than most of the cases in which it is true. In fact, the main point of this work is to draw attention to precisely those cases. More on this below.

The second point to clarify is that I do not mean that ‘free will’ discussions are really about *this* or *that* theory of moral responsibility. The proposal I make with regard to the dialectical situation of the ‘free will debate’—eliminativism notwithstanding—is to define the term ‘free will’ operationally, as though it were a philosophers’ technical term from the beginning. As I will go on to explain in more detail, ‘free will’ should be understood to mean something like ‘whatever it is, if anything, in virtue of which people are appropriate subjects of moral responsibility’. To put it a little simplistically, the claim is that talking about whether people

have free will *just is* talking about whether they are morally responsible agents.

Note that ‘being morally responsible’, in the sense of ‘being an appropriate *subject of* moral responsibility’, does not automatically mean that such a person is deserving of praise or blame, or anything else. This will depend on whatever specific view of moral responsibility that you hold — something that my argument here is silent on. It simply means that the person has met whatever conditions or requirements that there are on being the kind of subject that is a possible bearer of these ‘responsibility predicates’ (if anything is).

Note further that, even if one thinks that there is no possible theory of responsibility that is adequate—i.e. one is a ‘moral responsibility skeptic’—then my theory accommodates this. In such a case, one is simply a ‘free will skeptic’. On that point, it is worth observing that many of the recent popular-audience books and articles concerned with the ‘science has shown free will to be . . .’ theme are actually concerned with the notion of *moral responsibility*, and the implications for our notions of responsibility that might be drawn from the emerging science of human behaviour.³

³ See, for example, the neuroscientist Michael Gazzaniga’s (2011) book.

After making the argument above, I do not in fact go on to build a theory of moral responsibility. Far too much has already been written about that topic. What I do instead is take up the question of what does *not* fit into the characterisation of ‘free will’ just outlined. I suggested that there are some cases where moral responsibility is not the issue, and these are more interesting for me in what follows, partly because they have received comparatively less attention. In such cases, it turns out that *agency* is the salient issue. This has not previously been recognised, for the most part, because of the structure of the ‘free will debate’, and especially because of its fixation on the notions of determinism and indeterminism.

By making explicit which parts of the ‘free will’ issue are strictly about moral responsibility, we can group them together and address them in the appropriate theoretical context. By doing so, we also see that some of the ‘problems for our free will’ have nothing to do with moral responsibility — but they do directly bear on certain questions about the structure of agency. Apart from clarity and consistency, one of the advantages of this approach is that we can address those issues (i.e. those which are about agency) in their proper context: namely, in the light of work in the

philosophy of mind and action, and also recognise that issues pertaining to metaphysical determinism are not so relevant, something which goes against the prevailing view of what is important in the free will problem.

1.2 Overview of the Argument

In Chapter 2, I make the ‘second order’ argument in more detail, and claim that we could in principle do without the term ‘free will’ altogether. However, for practical purposes, I suggest we retain the term, but stipulate its meaning in the way already indicated. As well as the practical benefits just mentioned, in this chapter I make the claim that much of the existing literature on ‘free will’ *already* uses the term in this way, albeit implicitly.

I canvass a range of the most popular positions on the ‘free will’ issue, and show how they are ultimately concerned with the possibility of moral responsibility when they are writing about ‘free will’. I then turn to consider a possible counterexample to my view: namely the ‘semicompatibilism’ of John Fischer. The putative objection is that Fischer’s view doesn’t conform to the way of characterising the ‘free will problem’ that I suggest in this chapter, because he appears to be saying that ‘free will is not required for

moral responsibility'. However, my response to this is to suggest that, by contrast, we can see Fischer's view as a positive *demonstration* of the utility of my view, by slightly reframing his discussion of 'guidance control' and 'regulative control'. The putative objection thus turns out to serve as a case study that supports my argument.

In Chapter 3, I move away from the issue of moral responsibility, and pick up on the remaining cases in which moral responsibility is not the salient issue. In these case, as I have already suggested, we find issues that directly concern agency. In particular, a central feature of agency that I call '*agential control*', or just 'control' for short. On my view, agents necessarily control their actions in this way, and thus agential control is a *constitutive* feature of agency. In some ways, it might be better to think of 'agential control' as simply *being* my theory of agency, although I do not make an argument for that identification here. It is enough to accept that agential control is an essential feature of agency. In particular, I consider two examples from the 'free will' literature in which the notion of agential control is at issue. Traditionally, of course, this has not been recognised, since the discussions have been carried out as though they were about 'free will' and not agency.

The first case concerns the notion of ‘sourcehood’ or ‘authorship’. It is sometimes claimed that we could not be the ‘ultimate source’ of our actions unless we have ‘free will’ of whatever kind is being promoted by that philosopher: Randolph Clarke, for example, notes that by ‘acting freely’ we are ‘the ultimate source of [our] behaviour’ and that the action may be attributable to us as its ‘author’. Traditionally, it has been thought that it is because of ‘free will’ that we are the source or author in this way, and that threats to such sourcehood are threats to free will (and vice versa). Instead, I argue that the best sense we can make of these notions is as central features of *agency* — albeit with some important revisions. I make this argument in the context of Derk Pereboom’s ‘Four-Case Argument’ against ‘compatibilist free will’.

The second case concerns the notion of ‘agent causation’. In the literature on ‘free will’ there is a type of argument that turns up in the debate between those who think ‘free will’ is compatible with determinism, and those who do not. My argument here is that an important step in that argument depends on an assumption that cannot be sustained when it is recognised that *agency* is the salient notion at work. Hence, making the distinction outlined in Chapter 2 actually has implications for an important argument in the ‘free

will' literature. I conclude with some general remarks about the structure of agency.

In Chapter 4, I turn to consider the notion of 'agency' in more detail. Whereas in Chapter 3 I suggest that agency has something to do with 'causal integration', here I go on to make that idea more explicit. To begin with, I survey some existing literature to try and get a handle on what the concept of agency, as it stands, actually looks like. I suggest that the concept has been put to a wide range of uses, and thus we need some general constraints on what an acceptable notion of agency (for present purposes) is going to look like, that will enable us to narrow down the possibilities. To that end, I consider Helen Steward's view of agency as a useful starting point for inquiry.

I take two important points from Steward's work. First, I use her discussion as a springboard for considering the relation between folk psychology and agency, and more generally between cognitive science and metaphysics. Secondly, I pick up on certain features of her concept of agency, although I set aside others: in particular, the intriguing notion of 'being a centre of subjectivity', is a feature of Steward's concept of agency that she names but does not go on to explore in detail (due to the fact that her main focus is

elsewhere). Thus, making no claims about what Steward would say about my account of this notion, I go on to develop the idea of ‘being a centre of subjectivity’ in my own theory of agency. The upshot of Chapter 4 is that agency involves two central, and interconnected, notions that require further elaboration. These are *intentional realism* and *phenomenal consciousness*. I consider these in the following two chapters.

In Chapter 5, I consider the first of these notions, and unpack the connections between the concept of agency and that of intentional realism. I take the notion of ‘intentional realism’ to be fundamentally a claim about how we should understand human behaviour. It is associated with the ideas of ‘folk psychology’ and the so-called ‘belief-desire psychology’, because of its emphasis on the common sense notions of ‘belief’ and ‘desire’. Jerry Fodor, whose work I will briefly touch on below, has been a vocal supporter of some form of intentional realism.

For all the emphasis on the locutions of ‘belief’ and ‘desire’, however, the thesis of *intentional realism* is not strictly committed to those terms. Indeed, as Fodor emphasises, “the identity conditions for belief-states are vague and pragmatic in practise; perhaps

ineliminably so.”⁴ What is important, in this project especially, is that intentional realism is committed to the view that a true account of human behaviour cannot do without *intentionality* (i.e. the terms of a ‘mature scientific psychology’ must be intentional), and that such intentional states are *causally efficacious* — i.e. they figure in true causal explanations of human behaviour. Whether or not the specific concepts of ‘belief’ and ‘desire’ themselves figure in the final analysis is of less importance.

I consider two primary challenges to the notion of intentional realism, those of *eliminativism* and *instrumentalism*. Although the problem of eliminativism is fairly quickly dispatched, the challenge presented by instrumentalism leads into a deeper discussion of causation and its role in our concept of agency. Thus the discussion of this chapter ends up turning to the interventionist view of causation as offering a way of responding to the instrumentalist — but in a way that centrally involves the contribution of phenomenal consciousness.

In Chapter 6, then, I turn to the notion of phenomenal consciousness and its role in agency. It was suggested in Chapter 5 that it is a condition of intentional realism that there actually *are* the

⁴ Fodor (1992: 175).

right kind of interventions on our mental states, and not simply the theoretical possibility of such interventions. In this chapter I suggest that consciousness actually does function as an intervention in this way, and therefore that phenomenal consciousness becomes an integral part of the notion of agency.

But showing merely that an agent's intentional states can be affected by consciousness (via conscious perception) is by itself insufficient — it needs to be shown that such states *make a difference to behaviour*, since otherwise there is an important sense in which such states are epiphenomenal. It is a central requirement of Chapter 6 to defend my view against the challenge of epiphenomenalism.

To that end, I consider a recent challenge of this kind that has been called the problem of 'Zombie Action'. According to this problem, consciousness is not, in fact, constitutively involved in the production of action. And this is precisely what my view of agency requires: the purpose of Chapter 6 therefore is to defend a thesis that has been called 'Experience-Based Control'. That is, the thesis that phenomenal consciousness is involved in the production and control of action.

Finally, Chapter 7 wraps things up with some concluding remarks, and brings together the various strands of argumentation that have been considered thus far. I consider the claims that I have made about agency in the light of the original dialectical situation of the ‘free will problem’, and also note that the theory of agency provided is only a *sketch* of a theory, and that much work would need to be done to make the theory convincing as an account of agency in its own right. However, what it *does* show is the potential avenues that are opened up by recognising the role that the notion of agential control plays in the existing ‘free will’ debate, and by considering it in its proper context, divorced from the constraints of the determinism-indeterminism-compatibilism backdrop.

2

The Problem of ‘Free Will’

2.1 Introduction

It has been observed that ‘the free will problem’ is really a name for a cluster of related problems or worries.¹ Such problems include certain issues in the philosophy of mind;² concerns about physical determinism, reductionism, and mechanism;³ inquiry about the various ‘reactive attitudes’ that belong to our involvement and participation with others in human relationships;⁴ whether agents are ‘truly’ deserving of blame (and praise); and the justification of

¹ As Dennett puts it, ‘the free will problem’ is really a name for several *related* problems that are “tied together by a name and lots of attendant anxiety” (1981: 286).

² In particular, the work that begins with Libet’s work on conscious willing (1985); see also Wegner (2002).

³ See for example, van Inwagen (1983) or Honderich (1988) for classic statements of the concern with determinism; see Dennett (1973: 157-84) for a discussion of mechanism — but of course, also see La Mettrie (1996 [1747]).

⁴ Strawson (1962).

specifically *retributivist* punishment (i.e. over and above that needed for rehabilitation).⁵

In this chapter, I suggest that we need to rethink our use of terms like ‘free will’ and ‘the problem of free will’. As things are now, there is more heat than light in the literature on free will and moral responsibility. I am going to argue that we could eliminate the term ‘free will’ (and its various cognates⁶) from the literature altogether without significant loss to the substantive philosophical issues. While theoretically possible, however, it may be that wholesale elimination of the term proves impractical or undesirable, for whatever reason: in that case, I suggest that we move forward by *stipulating* the meaning of the term ‘free will’.

In fact, it is surprising that anyone believes that ‘free will’ is anything other than a philosopher’s term of art at this point: but the

⁵ See for example the debate between consequentialist views of justice and *desert based* views. See (Feldman and Skow 2015) for an overview. Often, the discussion over whether desert-based views of justice (esp. punishment) are appropriate depends on one’s view of ‘free will’. Libertarians tend to favour it, compatibilists don’t. In my view, this difference really turns on one’s theory of MR.

⁶ Here and in the remainder of the text, when I write ‘free will’, meaning to refer to the term itself, I mean to include all of the variations such as ‘free action’, ‘free agent’, ‘free’ and any other ways that the relevant sense of ‘freedom’ is supposed to modify other terms. It is this notion of ‘free’ that is important.

proliferation of work concerned with what the ‘folk’ think about the term suggests otherwise.⁷ It is supposed that we are constrained by the ‘folk concept’ of free will when working with the term, such that we are in danger of failing to talk about free will at all if we stray too far from that folk conception. Now, there may be some relatively common understanding of the term ‘free will’ among people without higher degrees in philosophy (the ‘folk’?) — although I doubt it. It doesn’t matter, because the terminological prescription I am making here explicitly requires that we use ‘free will’ as a technical term, with the meaning of that term stipulated in the way that I will now go on to explain.

I propose that ‘free will’ be understood along the lines of ‘the conditions necessary for moral responsibility (whatever they are)’. So, for example, when we say that an agent ‘has free will’, what this means (according to my view) is that she has met whatever conditions are necessary for her to be morally responsible for her actions. Acting ‘with free will’ means acting in such a way that those conditions are satisfied, and a ‘free action’ is one that can be appraised morally — i.e. it is an action performed by an agent who, at the time of acting, satisfied the various conditions for moral

⁷ For example, (Nahmias et al. 2005); Nichols (2004).

responsibility (whatever they are). I do not take a stand here on what those particular conditions might be, although examples from the literature include: having a ‘reasons-responsive mechanism’, the ability to respond to specifically *moral* reasons, or having an appropriate set of second-order desires.

The utility of defining ‘free will’ in this way will be demonstrated by the clarity that can be brought to various philosophical issues: indeed, as I will argue, most of the important work on free will is *already* concerned with moral responsibility, in more or less direct ways. Hence, there would be no loss of any substantial issues by simply defining ‘free will’ in this way. The obvious advantage is freedom from the task of staying faithful to a supposed folk usage of the term (what it ‘really means’), and uniformity across the philosophical landscape (i.e. no talking past one another because of a slightly different understanding of the term ‘free will’).

Finally, and most importantly for my purposes here, doing so will reveal a set of issues that are really *distinct* from the question of moral responsibility, but which have been run together with the above concerns, all coming under the heading of ‘the free will

problem’. Hence, the lack of consistency with the term ‘free will’ obscures a distinction that needs to be recognised.

Below, I will survey a range of examples from the literature to support the claim that moral responsibility is already the main driving force in many of the worries about free will. Simply put, if you think that free will matters because the possibility of moral responsibility somehow hangs on the question ‘whether we have free will’, then you’re interested in the conditions necessary for moral responsibility — so it would be preferable to go on with the clear and unambiguous project of determining what those conditions might be, without worrying about whether those conditions are *really* ‘free will’.

Objection: You’ve got it the wrong way around. Yes, we are interested in the conditions necessary for moral responsibility — but one of those conditions, perhaps the most important one, is *that we have free will*.

Reply: I would simply point out that ‘having free will’ must amount to *something* that is describable using terms other than ‘free will’ (or else it’s circular). We must have free will in virtue of something about the way we are, or about the world (or both). Hence, we could even drop the term ‘free will’ altogether and have

a discussion about whether we in fact have this thing (X) — and, perhaps more importantly, about whether it really is X that grounds moral responsibility, and not Y or Z — without any significant loss of substance. For the sake of retaining the term, however, we can simply *stipulate* that ‘free will’ is a way of referring to those properties or conditions (X, Y, Z, . . .), whatever they might be. When we ask, ‘What is free will?’, we are thus asking, ‘What are the conditions necessary for an agent to be morally responsible for her actions?’

The only further objection that could be introduced against my claim here, which I mentioned above, is that I have neglected the question whether X (or Y or Z) is *really* what ‘free will’ means: whether, for example, ‘X’ is what the folk have in mind when they talk about ‘free will’ or ‘freedom’. But the pressure to remain faithful to the folk usage of the term ‘free will’ is not a substantive metaphysical issue — it is not an issue that has anything to do with moral responsibility itself. It is a basically semantic question.

I do in fact think that such conceptual issues are important and interesting: in Chapter 4, I address the folk psychology of our concept of *agency*, for example. However, this is not what matters when one is engaged in the substantive project of articulating the

conditions necessary for agents to be actually morally responsible for their actions (or whether it is even possible to meet such conditions, or whether there even is a consistent set of such conditions to be articulated, etc.). These are substantive metaphysical, and empirical, issues. These are the issues that can continue to be addressed whether we retain the term ‘free will’ to describe them or not.

If the ever growing body of work in experimental philosophy did happen to turn out definitive evidence that there is a clear and precise sense of the term ‘free will’ that is ‘the’ folk meaning of the term, and hence is what the term ‘really means’, even this would not be a significant problem for my project here: I would simply give up the term ‘free will’, and continue with the two substantial metaphysical and empirical projects already discussed — that of articulating the conditions necessary for moral responsibility, and the other, distinct issue that I will shortly discuss.

2.2 Free Will as a Condition of Moral Responsibility

The first task, in this section, is to back up the claim that most of the discussion of ‘free will’ is *already* concerned with moral responsibility. I will do this by briefly surveying some

representative examples from the most popular positions that have been taken on ‘the free will problem’, and show how they are concerned with moral responsibility.

Indeed, it is even worth noting that the so-called ‘free will skeptics’ that are popular in the contemporary philosophical and mainstream literature generally turn out to be concerned with skepticism about moral responsibility. The three types of view I consider below cut across the main positions that have been taken on the problem, skepticism notwithstanding. These are *incompatibilism*, *compatibilism*, and *revisionism*.

2.2.1 Incompatibilism

Pereboom, for example, is a leading proponent of the suggestion that we do not have the free will required for moral responsibility: his ‘hard incompatibilist’ view is that while there *is* in fact a consistent description of the conditions under which we would have free will (conditions that he takes to be incompatible with determinism), these conditions are not instantiated in our world whether it turns out to be deterministic or not (because they

involve certain ‘agent causal’ metaphysical requirements that he believes we have good evidence against).⁸

The connection here between ‘free will’ and the conditions for being a morally responsible agent is explicit. For example, he writes that “the hard incompatibilist disavows freedom of the sort required for moral responsibility”, and that hard incompatibilism “is the view that there is no freedom of the sort required for moral responsibility.”⁹ The first paragraph of the book introduces the topic with a discussion of criminal behaviour and raises “the possibility that [such behaviour] may be caused by influences in upbringing or by abnormal features of the brain”, suggesting that this should make us doubt that such agents are morally responsible, before going on to claim that the main hard incompatibilist thesis is that, in some important sense, *all* our actions are like this.¹⁰ Free will is important, even for those who deny that we have it, because it is required for moral responsibility.

On the ‘optimistic’ side of incompatibilism—the libertarian view, according to which we have free will, and it is incompatible with determinism—Kane points out that it is possible to

⁸ Pereboom (2001) is the classic discussion. Also see more recently (2014).

⁹ Pereboom (2001: xxii, xxiii).

¹⁰ *Ibid.* (xiii).

understand what the free will problem is by considering moral responsibility: he notes that “Free will in the sense just described is also intimately related to notions of accountability, blameworthiness, and praiseworthiness for actions.”¹¹

Similarly, both van Inwagen and O’Connor apparently define ‘free will’ in terms of a specific *ability thesis*, namely, the ability to do otherwise than one actually does, for some particular choice or action. (And both philosophers give incompatibilist analyses of this ability).¹² Both philosophers take this ability to be necessary for moral responsibility — a common view among those who believe moral responsibility is not compatible with determinism, as well as some compatibilists who interpret the ability such that it is compatible with determinism. The ability to do otherwise—sometimes referred to as ‘the *freedom* to do otherwise’—is sometimes taken to simply *be* free will. When the ability to do otherwise is said to be required for genuine moral responsibility, this is once again a discussion of why *freedom*, or free will, is required for responsibility.

¹¹ See Kane (2002: 4), which is intended as an introduction to the free will problem. See also his (1996) for Kane’s own statement of libertarianism and associated discussion of the ‘significance’ of free will.

¹² Van Inwagen (1983) and O’Connor (2000).

In cases like this, my prescription for the term ‘free will’ will not change matters substantively: the problem is already framed in terms of a feature of agency which some philosophers argue is necessary for moral responsibility. The substantive debate between compatibilists and incompatibilists can then proceed via a discussion of whether the ‘ability to do otherwise’ is something that is or is not compatible with determinism, it being the thing that is required for moral responsibility.

In fact, van Inwagen is more certain about moral responsibility than he is about the incompatibility of the ability just mentioned: taking the fact of moral responsibility to be ‘a datum’, he notes that if science turned up proof that the world was in fact deterministic, he would then—and only then—become a compatibilist.¹³ Which is to say that, if determinism turned out to be true, this would show that free will must be compatible with determinism, because free will is a necessary condition for moral responsibility (and moral responsibility is, by hypothesis, established). Whatever one makes of the ‘datum’ claim, the connection between ‘free will’ and moral responsibility here is quite clear.

¹³ van Inwagen (1983: 223).

2.2.2 Compatibilism and Revisionism

Compatibilism is the view that free will is compatible with determinism. In fact, compatibilists often talk about ‘free action’, to emphasise that the conditions which are relevant to establishing whether an action was free are things like absence of constraint, or whether the action was performed under duress, and so forth. The reason seems to be that talk of ‘free *will*’ implies a strange *faculty* of the mind called ‘the will’, which would presumably have unusual metaphysical properties (such as being metaphysically indeterministic).¹⁴

The best way of making sense of a purported distinction between ‘free action’ and ‘free will’ that I can see is by way of a variation on Locke’s classic example of the sleeping man.¹⁵ Imagine

¹⁴ ‘Metaphysically indeterministic’ just means genuinely, robustly indeterministic. That is, where the very nature of the phenomenon involved is not deterministic, as opposed to the indeterminism being a function of a particular, limited, epistemic vantage point.

¹⁵ The original can be found in *An Essay Concerning Human Understanding*. (Book 2, Chapter 21, Section 12). The variation was invented by Kevin Timpe, in his Internet Encyclopaedia of Philosophy article “Free Will”, Section 1. (Last accessed 7/3/16).

that there is a woman, Allison, who is wondering whether or not to walk the dog. Before she makes a decision, she takes a nap. While she is asleep, a blizzard moves through the area and makes it physically impossible for her to leave the house. Upon waking, she decides to walk the dog. One might now use the distinction to say that Allison had free will (in making the decision to walk the dog), but not freedom of action (she could not in fact carry out the action).¹⁶

The *classical* compatibilist (Hobbes, Hume) view is that all there is to freedom is ‘freedom of action’, because all that matters is that one *can* do what one *decides* to do. Allison is not free to walk the dog, because there is a blizzard. In this context, insisting on the notion of ‘free will’ amounts to the claim that there is more to the question of whether Allison is free than whether there is a blizzard (for example). And it is at this point that the discussion of indeterminism usually begins.

¹⁶ On my view, this distinction is basically redundant. Is Allison ‘free’ to walk the dog? This can mean either of two things. (1) Is she morally responsible for walking / not walking the dog? Here it is plausibly relevant that the action is made physically impossible, but it depends on your theory of MR. (2) Is Allison, qua agent, capable of walking the dog? Yes, but she cannot exercise that capacity in this particular situation.

Compatibilists and incompatibilists can in fact agree on the claim that ‘freedom requires the ability to do otherwise’: they will simply disagree about the *metaphysics* of the abilities involved. Incompatibilists typically claim that such an ability is not compatible with determinism, because it requires that the agent have the ability to do otherwise, *holding fixed the laws of nature and the actual past right up until the moment of choice or action* — something that cannot happen if determinism is true. On the other hand, some philosophers¹⁷ have interpreted these abilities in a way that does not require determinism to be false. Importantly, however ‘ability’ is characterised here, both parties can agree that the ability to do otherwise is required for moral responsibility.

Finally, consider recent *revisionist* answers to the ‘free will’ issue. There is an argument, due to Vargas, designed to support the case for a revised conception of ‘free will’, which challenges the incompatibilist’s claim that the freedom required for moral responsibility should be construed incompatibilistically.¹⁸ The argument is roughly this: we hold people morally responsible, and, on the basis of certain moral judgments (e.g. acting wrongly), we

¹⁷ Vihvelin (2013) and Nelkin (2011).

¹⁸ See Vargas (2009: 51-2).

take steps to inflict punishment on these people — for example, through the prison system, or even through the use of capital punishment. The justification for inflicting this punishment is that they are *morally responsible* for a certain act; and to be morally responsible, as we have seen, is to meet certain freedom requirements.

Given that the incompatibilist construes these requirements in a way that makes them dependent on the falsity of determinism —*and* on the truth of various other metaphysical claims, regarding the location and role of indeterminism in the brain, for example— and given that we do not have good empirical evidence that such conditions are actually met (so the argument goes), we are in the troubling position of meting out punishment, along with moralised praise and blame, without strong evidence that people have met these requirements. Vargas writes:

A concrete example may make this point clearer. Consider *Fiery*. Fiery is a skeptical subject of a significant moral blame, and likely, punishment. Perhaps she faces the death penalty, if that is permissible, and if not, then some very significant censure where that variety or some large quantum of that

censure (whether blame or punishment) depends on the presumption of her being a libertarian agent. Now, let us imagine that Fiery demands to know why such treatment is justified.

Her libertarian persecutor must acknowledge that we have no evidence to support the hope that underwrites our treatment of her—that is, the hope that Fiery is, indeed, a libertarian agent. But Fiery will surely protest: the mere *possibility* that she deserves some extra quantum of blame or punishment beyond that required for say, rehabilitation, does not, by itself, make such treatment justified. After all, Fiery insists, there is also a chance that she—and everyone else—might not be libertarian agents. Indeed, this strikes her (especially now!) as considerably more plausible than her prosecutor’s insistence that libertarianism is true.¹⁹

Vargas thus concludes that if the empirical evidence does not ultimately come out in favour of libertarianism, then it would turn out to be “grossly unjust to hold her accountable to any degree

¹⁹ Vargas (2013: 7-8).

beyond the degree of blame and punishment warranted by non-libertarian considerations.”²⁰

He suggests that we should *revise* our ordinary concept of free will and eliminate the contentious commitments to indeterminism. He disagrees with the standard compatibilist that what we actually mean when we talk about free will in ordinary contexts is something that was never incompatible with determinism in the first place (we might just be confused about what we are actually committed to): rather, he acknowledges that our ordinary thinking does have certain incompatibilist ‘strands’, but that—in light of both (i) empirical plausibility, and (ii) the fact that we can get a workable conception of moral responsibility by invoking only features that happen to be compatible with determinism—we should just revise our ordinary concepts and eliminate any lingering incompatibilist commitments.

He thus claims that our concept of ‘free will’ is *mostly* as the compatibilist has it, but contends that there are parts of our conception of ‘free will’ which nonetheless remain committed to incompatibilism. Thus, he thinks that our existing conception is not

²⁰ Vargas (2013: 8).

wholly consistent, and for that reason we should revise it — in favour of compatibilism.

What is interesting about Vargas's approach is that he is guided in his revisionism by the constraints of developing what we might call a 'normatively adequate' conception of free will. That is, when deciding how to revise the concept, he suggests that we ought to take into consideration the implications for our practises of moral responsibility and justice, in particular, the practise of blame and punishment. As the above argument suggests, he believes that we should revise the concept so that it is compatible with determinism, because he believes this position is better supported by what empirical evidence we do have regarding human agency, as well as the thought that holding on to the incompatibilist conditions could have troubling consequences for our systems of justice, and more generally, our practices of blame and punishment.

In some ways, my own proposal here is similar to Vargas's revisionism about the compatibility / incompatibility issue, except that I am concerned with our use of the term 'free will' and its connection to both moral responsibility and agency. As I briefly suggested above, if there is a folk conception of the term 'free will'

it is—as Vargas believes—probably an inconsistent one. That is, the (folk) meaning of ‘free will’ has strands that are tied up with moral responsibility and the conditions for praise and blame, but also has strands which have little to do with moral responsibility at all, and are simply about features of agency. So the overarching prescription with which I began this chapter is a revisionary one: revise the concept of free will in favour of moral responsibility. The case for this revision is, like Vargas’s, a pragmatic one: *most of the time* this is how it is already used.

Furthermore, and most importantly for my own view, the benefits of this revision will be seen in the clarity that it brings on the *rest of the cases*, i.e. the aspects of our existing folk conception that are *not* primarily about moral responsibility. After this chapter, I have little more to say about moral responsibility: what is interesting to me is what is left over when those issues are set aside — that is, certain issues about agency that have, I claim, become obscured by their entanglement in ‘the free will debate’.

2.3 Moving Beyond ‘Free Will’

The foregoing discussion has been fairly abstract. In the remainder of this chapter I will attempt to make things more concrete by

considering a potential objection to my way of classifying the free will issue. By showing how this putative counterexample fails, and is in fact a good example of the utility of my own view, I will have made the issues more tangible by demonstrating an existing ‘case study’ in my way of framing the free will issue.

2.3.1 Objections to my view

What this very brief survey of the literature indicates is that, across a wide range of views about ‘the free will problem’, including views that are straightforwardly opposed to each other, there is a common thread: the view that free will is in some way implicated in our concept of a morally responsible agent. Specifically, it is taken to be a metaphysical *condition* for the correct attribution of moral responsibility to some agent (for some action). There may also be epistemic and situational conditions as well, although I will not consider them here, as they do not change the substance of my claim — I am not concerned with the *actual* conditions of moral responsibility at all, but rather am pitching my argument at one level of abstraction from that debate, as it were.

Now, libertarians suppose that we in fact meet this metaphysical condition, and that it has incompatibilist satisfaction

requirements. Hard incompatibilists such as Pereboom agree that meeting this condition requires the falsity of determinism — they simply disagree that we meet it. Thus, on this way of putting things, the hard incompatibilist says that we ‘lack the free will required for moral responsibility’, i.e. we fail to meet a metaphysical condition that is necessary for being morally responsible agents.

Many compatibilists also share the view that there is a certain metaphysical condition—i.e. having free will—which is necessary for us to be morally responsible: but they hold that meeting that condition is perfectly compatible with the truth (or indeed the falsity) of determinism. Often this type of view is put in terms of the ‘ability thesis’ indicated above, where ‘free will’ is characterised as the ability to do otherwise, and that ability is interpreted as being compatible with determinism. As we saw there, the disagreement between this type of compatibilist and the incompatibilist is not about whether agents need to meet this metaphysical condition, but rather it is about *whether* meeting that condition is compatible with the truth of determinism, and this turns on the particular analysis of abilities that is given in each case.

For example, consider the recent work by the ‘new dispositionalists’, which is aimed at providing a theory of abilities that is compatible with determinism, such that the compatibilist can agree (*pace* Frankfurt-style objections) that the agent must have the ability to do otherwise, without requiring the falsity of determinism.²¹

Finally, the revisionism developed by Vargas, for example, ends up in a position similar to the type of compatibilism just indicated (e.g. claims that there are metaphysical conditions on being morally responsible, and they are compatible with determinism), but does *not* claim that this is an accurate reflection of our ‘folk’ conception of free will. That is, unlike the compatibilist, who effectively claims that everything we ordinarily believe to be required for free will and moral responsibility in fact *turns out* to be compatible with the truth of determinism, the revisionist just accepts that we probably do have incompatibilist intuitions about these matters. They simply claim that intuitions must give way to theory. The revisionist goes on to ‘prune’ the

²¹ See, for example, the recent monograph by Kadri Vihvelin (2013). For a discussion of the ‘new dispositionalists’, and the origin of that label, see Clarke (2009).

concept of free will by eliminating those incompatibilist intuitions, guided explicitly by the demand for a normatively adequate account of the ‘freedom’ that is required to justify our attributions of moral responsibility.

I have suggested that ‘free will’ is usually understood as a metaphysical condition that is required for the possibility of moral responsibility (more precisely: for an agent to be morally responsible for an action).²² My claim is that we should drop the term ‘free will’ and straightforwardly investigate what these conditions are. This simplifies the theoretical constraints on the theory: first, determine what the most adequate theory of moral responsibility is, and then whatever conditions (X) turn out to be necessary for ‘being morally responsible’ on that view are what we investigate when we are interested in whether some action is ‘free’. There is no need to look elsewhere for clues as to what ‘free will’ might be.

One possible objection to this way of characterising things might begin by pointing to the so-called ‘semicompatibilism’

²² Other than actions, agents may be thought to be potentially morally responsible for *omissions* or *outcomes*. I will simply discuss the case of *actions* in what follows because I do not believe the present issues are significantly affected by the differences between such cases as actions or omissions, etc.

developed by John Fischer, and sometimes in conjunction with Mark Ravizza: this form of compatibilism is notably different from the kind indicated above (the compatibilist-abilities view), because it denies that free will—conceived as the ability to do otherwise—is necessary for moral responsibility. Thus, Fischer agrees with the incompatibilists that this ability is not compatible with determinism; but at the same time, he claims that moral responsibility *is* compatible with determinism, because the facts that are relevant to this question are only facts about the ‘actual sequence’, and not what alternative possibilities there are.²³

At first, the claim that ‘free will is not required for moral responsibility’ seems to fly directly in the face of the argument I have been making here, part of which was the suggestion that most of the literature already takes ‘free will’ to mean something like ‘the conditions required for moral responsibility’. Now it seems that a prominent philosopher writing about these issues holds exactly the opposite view.

However, this is not the case. Firstly, of course, I could simply emphasise the revisionary nature of my argument and say ‘so much the worse for Fischer’s view’. Indeed, I only claimed that

²³ See Fischer (1994) and the essays collected in (2006); Fischer and Ravizza (1998).

‘most’ of the literature already made the connection, not that it was *unanimously* held. This would be a valid response, but in fact there is a stronger claim to make: I suggest that, on the contrary, Fischer’s view is basically an example of my own suggestion put into practise, albeit not explicitly under that rubric. Hence, examining the Fischer case will both deflect this putative objection to my view, and at the same time help the positive case for it.

2.3.2 Fischer and Semicompatibilism

According to semicompatibilism, moral responsibility is compatible with determinism (as well as indeterminism): as Fischer puts it, “Our fundamental nature as free, morally responsible agents should not depend on whether the pertinent regularities identified by the physicists have associated with them (objective) probabilities of 100 percent (causal determinism) or, say, 98 percent (causal indeterminism).”²⁴ Fischer believes that the possibility of moral responsibility should not ‘hang by a thread’, awaiting the deliverances of theoretical physics. At the same time, he believes that the *freedom to do otherwise* is not compatible with determinism, pointing to the variations on the ‘consequence argument’

²⁴ Fischer (2006: 5).

developed by incompatibilists such as van Inwagen.²⁵ Thus ‘semicompatibilism’ is, roughly, the view that moral responsibility is compatible with determinism, even though the freedom to do otherwise (as the incompatibilist traditionally characterises it) is not.

This way of putting things is slightly unorthodox, and so it does not straightforwardly fit the characterisation I have given above. I have pointed out the way in which much contemporary work views ‘free will’ as a metaphysical condition for the licensing of moral responsibility; but Fischer distinguishes what is often thought of as the primary ‘metaphysical’ characterisation of free will—that is, ‘the freedom to do otherwise’—as being *distinct* from the question of moral responsibility (by claiming that ‘free will’ is neither necessary nor sufficient for agents to be morally responsible). So it might seem as though my attempt to eliminate the term ‘free will’ puts me at risk of conflating issues that should not be run together in this way.

In fact, things become much *clearer* if we avoid using terms like ‘free will’ and ‘freedom’ to describe Fischer’s view (or any view). We *could* use the term ‘free will’ to mean ‘the ability to

²⁵ Fischer (1994) deals with these incompatibilist arguments.

otherwise’, which seems to be a fairly common move when articulating the view that morally responsible agents must have the ability to do otherwise (indeed, it is sometimes phrased as ‘the *freedom* to do otherwise’).

But on this way of speaking, we would be then forced to describe Fischer’s view as the claim that ‘free will is not required for moral responsibility’, which hardly makes the view clearer; nor does it do anything to help one understand what *is* required for moral responsibility on Fischer’s view.

Why not distinguish two kinds of freedom? These might correspond to the two types of *control* that Fischer outlines in his theory of responsibility: guidance control and regulative control. In his terms, ‘regulative control’ is the kind that affords the freedom to do otherwise that he believes is not compatible with determinism; it is only the less metaphysically demanding ‘guidance control’ that is required for moral responsibility, and this remains compatible with determinism because it is concerned only with ‘the actual sequence’, rather than what could or would happen under certain conditions.

But these moves add nothing of substance to the debate. Dropping the term ‘free will’ altogether for a moment, we might

say that Fischer's semicompatibilism is concerned with articulating the conditions that agents must meet to be morally responsible for their actions, and these conditions include various constraints on their actual situation (in this case, facts about their 'reasons-responsive mechanism', i.e. whether they are appropriately responsive to reasons, including moral reasons), and the 'ability to do otherwise' is *not* one of those conditions. Hence, if you had simply taken 'free will' to *be* 'the ability to do otherwise', you would now be able to say that Fischer believes that 'free will is not required for moral responsibility'.

Indeed, we should also avoid using the term 'freedom' to describe that which Fischer claims is incompatible with determinism: instead, we should say that he has identified one possible *feature of agency* (viz. the 'ability to do otherwise'), which turns out to be incompatible with determinism — but, according to semicompatibilism, this is not one of the features that is required for moral responsibility. If agents in the actual world do not possess this ability (for example, if determinism were true), this would not be a barrier to the legitimate attribution of moral responsibility to those agents, although it may have other implications for agency. And the traditional debate about whether agents can be morally

responsible for their actions, without access to ‘metaphysically open’ alternative possibilities (i.e. the ability to do otherwise) need not be significantly affected.

If we wish to retain the term ‘free will’, then clearly the distinctions just made can be maintained by stipulating the meaning of the term in the way that I’ve suggested above. For Fischer, then, ‘free will’ is guidance control, because it is guidance control that is necessary for moral responsibility. By contrast, regulative control is something that requires the ability to do otherwise: this is not free will, but it may or may not be a valuable feature of agency for some other reason.

The fact that it is possible to explain the substance of Fischer’s view without using the term ‘free will’ only supports the argument of the present chapter. As noted earlier, it is really an open question whether or not we should retain the term ‘free will’ at all: certainly it is not *necessary* to retain the term in order to continue any of the substantive discussions that I have been considering here.

Fischer’s semicompatibilism actually provides us with a good case study in the possibility of dropping the term (or at least stipulating a clear use for it). Firstly, Fischer does not claim that

‘free will’ is what’s necessary for moral responsibility, whether that is construed as compatible or incompatible with determinism: what is required, on his view, is an appropriately functioning reasons-responsive mechanism. Of course, one might disagree with this view about what is necessary and sufficient for an agent to be morally responsible for an action, but in this case it is clear where the disagreement lies. One might disagree with this view, insisting that a certain kind of ability to do otherwise really is required for genuine moral responsibility: but in this case, the disagreement is about which facts are relevant to an agent’s moral responsibility — facts only about the actual situation, or certain *modal* facts about the available possibilities at that time.²⁶

In fact, dropping the term ‘free will’ from this debate allows a more *fine-grained* discussion of the relevant issues. It may be that agents meet certain conditions that justify *one* aspect of moral

²⁶ Indeed, one could dispute the characterisation of ‘the actual sequence’ as being something that can be properly described without reference to any modal facts: in short, one might argue that in order to properly understand what *does* happen in a situation, one must understand certain facts about what *can* happen there. (See Steward 2009: 88 for more on the “entanglement of modality with the characterisation of actuality”.) My point here is just that this discussion could only be *clarified* by carrying on the argument without the use of ‘freedom’ terms.

responsibility—for example, a consequentialist approach to punishment—while at the same time failing to meet further conditions that would be required for other aspects such as moralised praise and blame, and genuine ‘desert’. In this way, the question whether agents are morally responsible is not an all-or-nothing question: to put it differently, the question whether agents meet ‘the’ condition necessary for moral responsibility (previously picked out by the term ‘free will’) is not the right question to ask, because there are several conditions which correspond to different kinds of (or features of) moral responsibility, and no one of these ought to be prioritised over the other.

Furthermore, it even remains possible to engage with the various incompatibilist ‘consequence arguments’, as Fischer does, without using the term ‘free will’ (or similar). The consequence argument describes an ability, which is usually described as the ability ‘to do otherwise’, or ‘choose otherwise’, and purports to show that this ability is incompatible with determinism. To obtain the further conclusion that ‘free will is incompatible with determinism’ it is of course necessary to add the premise that “‘free will” is “the ability to do otherwise, holding fixed the laws of nature and the past”’, or some similar premise, in order to make the

connection. As I suggested above, if one believes that this particular ability to do otherwise is actually a necessary condition for genuine moral responsibility, then the consequence argument is going to be very important anyway, even without the additional premise.

Assume that the consequence argument is sound: there is an ability, described in the consequence argument, that is indeed incompatible with determinism. What is important to recognise is that the ability currently being discussed—that is, ‘ability’ according to the traditional incompatibilist—is simply one among many related abilities that are relevant to dissuasions of agency, action, and moral responsibility. According to my view, the traditional incompatibilist position can therefore be redescribed as the view that *this* ability is necessary for moral responsibility, and that such an ability is not compatible with determinism.

One standard compatibilist response, which became popular after Frankfurt’s seminal paper, is to deny that the ability to do otherwise is necessary for moral responsibility at all: in the present terms, of course, we should say that *an* ability to do otherwise is not necessary, viz. the ability that the traditional incompatibilist has in mind when they talk about the ‘ability to do otherwise’ (and which is not compatible with determinism).

Indeed, Fischer's view is often described simply as the claim that 'the ability to do otherwise is not required for moral responsibility'.²⁷ And this interpretation has apparently been supported by Fischer's defence of various Frankfurt-style examples. But, as Christopher Franklin has argued, if we allow that there are multiple senses of 'ability', we can see that Fischer in fact only disputes the necessity of the particularly 'strong' kind of ability to do otherwise just indicated.²⁸ He nonetheless requires *some* kind of ability to do otherwise, or access to some kind of alternative possibilities, but not the kind usually picked out by standard incompatibilists (the 'strong' kind).

The so-called 'Frankfurt cases' are counterexamples to the Principle of Alternative Possibilities (PAP), which is simply the claim that we have already encountered above: that an agent must have the ability to do otherwise (thus having 'alternative possibilities' available) in order to be morally responsible. Frankfurt cases purport to show that there are cases in which an agent is intuitively responsible for some action A, but where they could not

²⁷ For example, Campbell (2011: 89).

²⁸ Franklin (2015).

have avoided doing A (or choosing to do A).²⁹ Hence the PAP cannot be right, according to the argument.

The Frankfurt case itself is really a *template* for constructing any number of similar cases: the details of any given case are not as important as the structure of the example. The basic idea is in fact very similar to Locke's 'sleeping man' example that we considered above. Here is a basic Frankfurt case: Jones is a Democrat, and always votes Democrat in the relevant elections. Black is some kind of interested third party who wants to *ensure* that Jones definitely votes Democrat in the coming election, so he inserts a device in Jones's brain in order to monitor the relevant decision process. If Jones should show any signs of voting *other* than Democrat, Black will activate a control mechanism in the device and thereby ensure that Jones votes Democrat. As it happens, Jones does not even consider an alternative, and votes Democrat anyway. Thus Black never needs to activate the device, and hence does not intervene in Jones's decision or action at all.

The intuitive response to this case is that Jones is responsible for his vote — even though he could not have done otherwise.

²⁹ Frankfurt's original paper is (1969). Fischer's overview of the Frankfurt cases can be found in Fischer (2010).

Fischer's way of putting this is to say that it's the 'actual sequence' that matters for responsibility, and this is part of his claim that an agent need not have the (strong) ability to do otherwise in order to be moral responsible. By contrast, the central feature of Fischer's account is that an agent is morally responsible only if they have an appropriately functioning reasons-responsive mechanism. And this is a matter of what occurs in the *actual* sequence, rather than in the realm of alternative possibilities.

Yet Franklin points out that, for Fischer, what it means for an agent's 'reasons-responsive mechanism' to be sufficiently sensitive to reasons, and for it to respond appropriately, must be characterised *in modal terms*. Whether an agent has a properly 'reasons-responsive' mechanism is evaluated by asking whether there is a possible world in which the mechanism *does* respond appropriately to reasons, i.e. by asking whether there is "*some* possible world in which there is a sufficient reason to do otherwise, [and] the agent's actual mechanism operates, and the agent does otherwise."³⁰

The only difference with regard to the usual discussions of 'alternative possibilities' in the literature is that *this* possible world

³⁰ Fischer (2006: 68).

need not be the ‘closest’ possible world. By contrast, the traditional incompatibilist holds a view of abilities (i.e. the ‘strong’ view) according to which we must hold fixed everything about the actual past apart from the choice or action in question (i.e. the choice or action that is supposed to be ‘free’) if we are to say whether the agent had the ability ‘to do otherwise’ — if determinism is true, of course, there are no such worlds.

Secondly, for Fischer, the *agent’s* responsibility is grounded in the ‘responsibility’ of the reasons-responsive mechanism: an instance of a general move that is common to reductionists about agency. In this case, the agent is responsible *in virtue of* the functioning of her reasons-responsive mechanism. This move, from mechanism to agent, is fairly uncontroversial unless one is a strict antireductionist about agency (e.g. one holds an ‘agent causal’ view). Thus the agent must have *some kind* of ability to do otherwise, because the reasons-responsive mechanism must have such an ability (or disposition).³¹

Yet, at no point is it helpful to pick out one of these abilities and privilege one of them with the label ‘free will’. In fact, the

³¹ For more on the connection between ‘ability’ and ‘disposition’ in this debate, refer to the ‘new dispositionalists’ cited above.

traditional characterisation of ‘free will’ as *the ability* to do otherwise might actually confuse the issue, with the implicit suggestion that there is only *one* ability in this area that is relevant to the question. The primary issue is what must be true about an agent if they are to be morally responsible for their actions. Even if one insisted on retaining the term ‘free will’ in this context, it would need to be defined *after* the investigation into agents’ abilities just noted: once it was determined which ability, or abilities, were required for moral responsibility, then one *could* stipulate that such an ability constituted free will — because ‘free will’ just refers to whatever conditions, if any, are required for moral responsibility (if there are no such conditions, then one is a ‘free will skeptic’, i.e. a skeptic about moral responsibility).

The further advantage of proceeding in this way is that one can sidestep debates about whether one has accurately characterised the ‘folk’ conception of ‘free will’ in the account. The revisionist approach offered by Vargas already goes some way towards this kind of view, but nonetheless there remains some pressure to avoid revising the term too far from the common ‘folk’ conception in order to be seen as a plausible revision of *that* term, and not simply the introduction of a new term. An ‘eliminative’

view does not have this limitation: since it does not use the term ‘free will’ at all, there is no obligation to demonstrate that one is remaining sufficiently faithful to the folk usage for the view to count as plausible.

Instead, such a view would have us discuss the metaphysical and ethical issues *directly*, by (i) investigating the metaphysical features of agency, and (ii) determining what must be true of an agent for her to be an appropriate subject of the different forms of moral responsibility — without concern for whether this description ‘deserves’ the name ‘free will’. If anyone suggests that one is not, after all, discussing *free will*, but something else altogether, one could simply ask them why they were interested in an analysis of free will in the first place, and refer them to the operational definition of ‘free will’ given above, which connects the term to the discussion of which conditions are necessary for morally responsible agency.

2.4 Conclusion

I suggest that Fischer’s semicompatibilism, far from being a problem for my view, can actually be seen—with a few small amendments—as a good example of the progress that can be made

on these issues without a reliance on the term ‘free will’, and instead with a focus on the conditions required for moral responsibility: which, in Fischer’s case, takes the form of a discussion about different forms of *control* that agents may have with regard to their actions. It would only be a step backwards to try and artificially constrain Fischer’s work with the language of the classical ‘free will’ debate.

The term ‘free will’ could in principle be dropped altogether, but for various practical purposes, it may be better to simply retain it and stipulate that it is a technical term. This stipulation involves making the operational move suggested above, and defining ‘free will’ in terms of whatever conditions or properties turn out to be necessary for morally responsible agency.

In the remainder of this work, I will have little more to say about moral responsibility (and hence about free will, as I understand it). What I will do next is look more closely at what substantive issues get ‘left over’ when we set aside the concerns about moral responsibility noted above — that is, issues which have so far become entangled with the ‘free will debate’ but which are not directly concerned with moral responsibility in the ways that I have considered above. *These* issues turn on some important

and interesting features of agency, and I will make the case for considering them independently of 'the free will debate' and of the related concerns about determinism and indeterminism which have become so essential to discussions of 'free will'.

3

Control and Sourcehood

3.1. Introduction

When we drop the term ‘free will’ from our philosophical vocabulary, and then look to find some issue, not involving the term ‘free will’ or its cognates, that the relevant parties to the debate substantially disagree over, we find that the most common disagreement is about the conditions necessary for moral responsibility. Put simply: most of the time, people worry about ‘free will’ because they worry about whether or not we can be morally responsible. I cited various sources of evidence for this claim in the previous chapter, ranging across incompatibilist and compatibilist views, and certain revisionist views about free will. The point was to illustrate the common connection between the use of the term ‘free will’ and concerns about moral responsibility across a range of the most popular positions that have been taken in the ‘free will debate’.

With that established, the question addressed here is whether there is anything more to ‘free will’ that does not relate to moral responsibility in this way. In fact, as I suggested in the previous chapter, we do find some claims in the literature that seem to attribute an importance to ‘free will’ (or ‘free action’) that does not directly follow from its connection to moral responsibility.

These issues concern a certain central feature of agency, which I am going to call ‘*agential control*’. The term ‘control’ already appears in the literature on free will and moral responsibility, but I am stipulating the present term. Rather than use the phrase ‘agential control’ each time, in order to mark this distinction, I will simply note here that, unless specified otherwise, my use of the term ‘control’ should be read as shorthand for ‘agential control’.

Agents necessarily control their actions in this way. Strictly, it would be more precise to say that it is the behaviour, or ‘bodily motion’, that is under the intentional control of the agent, and it is this *causing-of-behaviour-by-the-agent* which is an action. That is, when an agent’s behaviour is under her intentional control in the right way (to be spelled out later), this *constitutes* the agent carrying out an action — and we can say about this behaviour that ‘it was under her control’. I will often simply say that ‘the *action* was under

her control', or that she 'controlled the action'. This should be read as a convenient shorthand for the longer, but more precise, claim.

Reading these issues (i.e. 'free will' issues) as being about control is, of course, a stipulation of mine — but it is one that enables us to tie together several themes common in the literature, and to get a better perspective on what the problem really is. One of the reasons that these issues have not been addressed in this way is that the dialectic of 'free will' tends to obscure the distinction that I am making here.

This is because on anyone's account, 'free will' is something *had* by agents, not something equivalent to agency itself (agency is a more or less implicit precondition for free will). By contrast, control is a matter of the basic metaphysics of agency. So one of the basic 'conceptual housekeeping' tasks in this chapter is to point out that the terminology of 'free will' has obscured an important distinction between issues about moral responsibility and issues about the metaphysics of agency.

I will make these claims more concrete in the remainder of this chapter, but the basic claim can be quickly stated here. In the free will debate, there are two areas where this notion of control is doing unnoticed work in the background: in the notion of

‘sourcehood’ or ‘being the source of the action’, and in the woefully misunderstood notion of ‘agent causation’. These issues are about agency, and the control that agents have over their actions.¹ They are *not* primarily about moral responsibility, and hence are not about ‘free will’ in that sense at all — contrary to how they are usually presented.² The only sense in which they are ‘about’ free will—i.e., about moral responsibility—is the sense in which agency is a basic prerequisite for any action at all.

3.2 Sourcehood and Agent Causation

When the great importance of free will needs to be stressed it is usually the notion of moral responsibility that gets pressed into service. But sometimes an alternative explanation is given for its importance, and then terms such as ‘sourcehood’, ‘authorship’, or

¹ See the point above regarding this locution: “controlling their actions”. I frequently use this technically incorrect formulation for the purposes of clarity.

² Of course, I don’t favour the term ‘free will’ at all, but it’s open to someone to accept what I say here and argue that ‘free will’ means both things, or that we can define two kinds of ‘free will’ — the kind that is about moral responsibility, and the kind that is about agential control. Or to stipulate that ‘free will’ is about moral responsibility only, and the latter issue is simply about agency. If forced to choose, I would prefer to just do without the term ‘free will’ altogether.

‘ultimacy’ get brought out. For example, it might be claimed that a certain view of free will shows how we can be ‘the *ultimate source* of our actions’ in a way that we could not be without subscribing to that view of free will.

When considering the importance of free will, Randolph Clarke writes the following:

Indeed, even apart from the issue of moral responsibility, the attributability of actions and some of their consequences to free agents may be regarded as something of value. In acting freely, one is an *ultimate source of one’s behavior* and of its consequences, which may be attributable to one as their author.³

According to this view of free will,

free agents are (in an important respect) *originators of their actions* [...] a free action may be attributable to the agent in a

³ Clarke (2003: 7). My emphasis.

certain way, and some of its consequences may be so attributable⁴

It is the ‘certain way’ in which an action ‘gets attributed to the agent’ that is the focus of this chapter. The question whether an agent is the originator, or the ultimate source, of her actions is the question whether she *controls* those actions (agential control). Clarke sometimes calls this the ‘condition of production’.

A second case in which control is the salient issue is that of ‘agent causation’. The theory of agent causation has been wildly unpopular in the literature on free will, and has historically been associated with incompatibilism (also fairly unpopular). The basic idea here is that the causation that occurs when there is an action is different in some important respect from other kinds of causation in the world. Typically, the claim is that the agent is a *substance cause* of her action, whereas non-agential causes fall under the category

⁴ Clarke (2003: 6). My emphasis.

of ‘event causation’.^{5,6} It is the agent herself, as a substance, that is the cause of the action and not, for example, an event that *involves* the agent, or certain mental states *of* the agent.⁷

I will consider a specific instance of the agent causal account of action, as it features in an argument about free will. As I will go on to explain, there are two ‘types’ of incompatibilist about free will: those who invoke the agent causal theory, and those that accept a fairly standard event causal model of action, although both remain incompatibilists about free will. Now, there is a particular

⁵ As above, many of the more plausible versions of agent causation note that agents cause movements or changes in their bodies (or mental states), and it is that causally complex event that *is* an action — as opposed to the claim that agents as substances *cause their actions*.

⁶ This is only one account of the uniqueness of agent causes. Other, less plausible, accounts have it that the causal *relation* itself is somehow metaphysically special, or that agents are *uncaused causes*.

⁷ Some care should be taken here, because even if the agent is somehow a *sui generis* cause ‘of her action’, there still must be *some* event that occurs at that time, viz. the event of the agent’s agent-causing the action. Note that the view suggested above, on which it is a bodily motion that the agent directly causes, makes more sense of this: there *is* an event, and it is the agent’s ‘agent-causing’ the bodily motion — *this* event is an action. To preempt my later argument: this is why I suggest that it is causal *integration* that matters, and not *what it is* that does the causing, so to speak.

type of argument that challenges the coherence of the incompatibilist position in general, one that targets the reliance on indeterminism. This type of argument comes from the perspective of those who think free will is compatible with determinism, so the upshot is supposed to be that incompatibilism is incoherent in principle, and a compatibilist account is to be preferred.

There have been several particular instances of this argument, going under such informal names as ‘the luck objection’, the ‘chance objection’, or the ‘*Mind* argument’. Derk Pereboom calls it the ‘Humean Challenge’. For present purposes, we can group all of these together according to the general form that they share — criticism of the reliance on indeterminism as a necessary feature of the account of ‘free will’. I will call this collection of arguments the *coherence problem for incompatibilism*.

Here is the dialectical situation: *agent*-causal incompatibilists often defend their view against *event*-causal incompatibilists by claiming that the latter cannot respond adequately to the coherence problem. They claim that only by accepting the agent causal theory can we ensure that the agent has a special kind of ‘control’ (not agential control) over her non-deterministically caused actions, and it is (only) this kind of ‘control’ that is sufficient to resist the

coherence problem. Hence we can save incompatibilism from the coherence problem, but only by being agent causal incompatibilists.

The notion of agent causation thus becomes tied to a particular defence of incompatibilism, which is a theory about free will, and so it then seems that agent causation is something that has primarily to do with free will. Plainly, the notion of agent causation here is connected to an idea of ‘control’, and the very idea of *agent* causation is such that it already involves the notion of an agent in some sense. However, it is standardly addressed only within the context of the free will debate: it is more or less universally assumed that agent causation is an incompatibilist concept. With one or two recent exceptions, the theory of agent causation has historically been adopted only by incompatibilists about free will.

Now, these two issues are directly connected (sourcehood and agent causation), since both concern the way in which agents control their actions. I suggest that agent causation, like the notion of sourcehood, is really a particular kind of response to problems that threaten control (agential control). And both typically get presented as part of an incompatibilist theory of free will, which is a mistake: the notion of control is not partisan in this way, because

the notion of *agency* is not partisan — everyone, compatibilist and incompatibilist alike, needs an account of agency.

In what follows, I unpack both of these debates and argue that they turn on the notion of agential control — they concern the metaphysics of agency. The problem is that they have been co-opted into the free will debate, and it has been assumed that they are artefacts of a specific sort of incompatibilist view of ‘free will’. In fact, the issues should concern both compatibilists and incompatibilists about ‘free will’ (that is, about the conditions necessary for moral responsibility). This is because they are not *about* free will, but about agency.

One possible reason for this confusion is as follows. If you think that ‘sourcehood’ is part of free will (hence required for moral responsibility), you might think up cases in which the relevant notion of ‘sourcehood’ was absent, in an effort to show that agents in that condition could not be free (hence not morally responsible). But if ‘being the source of your actions’ actually picks up on features needed for *control*, then in the absence of those features there is going to be a problem for agency — and agency is necessary for moral responsibility, and of course for ‘free will’ (such as it is). So it might indeed be possible to evoke intuitions of non-

responsibility or ‘non-freedom’ in such cases, but not for the reason that is supposed.

There are at least three important consequences of recognising that the issues here are concerned with agency. Firstly, the issues are nonpartisan: both compatibilists and incompatibilists about free will need to have a theory of control. As a result, the agent causal incompatibilist loses an important dialectical move against event causal versions of incompatibilism.

Secondly, since these issues have been tied up with the particular motivations of the free will debate, they have not been exposed to the body of literature in the philosophy of mind and action which directly concerns such issues. It is likely that doing so will be illuminating, because there is much important work being done on agency in other contexts.

Thirdly, and most strikingly, since agent causation, sourcehood, ultimate responsibility, etc., are really just different ways of getting at the notion of control, it turns out that if certain incompatibilist arguments are correct, and ‘sourcehood’ or ‘agent causation’ really is not compatible with determinism, then *agency* is not compatible with determinism. I address the question whether we actually have good reason to think that agency is incompatible

with determinism in the next chapter, in the context of Helen Steward's views on agency. The task here is simply to point out that this is the logical consequence of recognising that these 'free will' terms (sourcehood, agent causation) are really picking up on basic features of agency, viz. control.

The argument proceeds as follows. First I address a recent view which has been called 'source incompatibilism', and which is placed in contrast to the more traditional 'leeway incompatibilism'. This new version of incompatibilist free will is notable for its claim that alternative possibilities are not what is necessary for moral responsibility, but rather it is whether or not the agent is the *ultimate source* of her actions that matters.

I consider one of the most influential arguments for this claim, due to Derk Pereboom, called the 'Four-Case Manipulation Argument'. This manipulation argument is designed to show that moral responsibility is incompatible with determinism because if determinism is true, then the agent is not in control of her actions, which are determined by *sources* outside of her in a responsibility-undermining way. Hence 'being the source of your actions' is not compatible with determinism, and free will (moral responsibility) requires us to be the source of our actions in this sense. Following a

recent response to Pereboom, I argue that the manipulation in the Four Case argument actually undermines *agency*, and this is what leads to the impression that moral responsibility is ruled out.

I then turn to the notion of ‘agent causation’ as it features in the defence of certain incompatibilist views. Such agent causal incompatibilists invoke the thesis of agent causation in order to defend their account against the coherence problem. According to the coherence problem, incompatibilism is incoherent or empty because the presence of indeterminism undermines the agent’s ‘control’ over their actions, or more generally, because it at least fails to *enhance* their ‘control’ vis-à-vis compatibilist accounts.

Contrary to this view, I suggest that agent causation is, quite rightly, a response to perceived difficulties with control, but that this control is a central feature of *agency* (agential control). It is not therefore something that is only of importance to those who believe ‘free will’ is incompatible with determinism, and are looking for a way to accommodate indeterminism into their account without leaving themselves open to the coherence problem. Hence the theory of agent causation is of concern to both compatibilists and incompatibilists about free will (something which has recently been recognised by a small number of authors, but for different reasons).

3.3 Source Incompatibilism

First, consider the related notions of ‘sourcehood’, ‘origination’, or ‘being the source’ of your actions. One place that these ideas show up is in discussion of a particular kind of incompatibilism about free will, which has been called ‘source incompatibilism’ (distinguished from so-called ‘leeway incompatibilism’).

The primary difference between these ‘types’ of incompatibilism is based on what it is that is supposed to be incompatible with determinism. In what is taken to be the more traditional type of incompatibilism—‘leeway’ incompatibilism—it is supposed that the agent must have a certain kind of *alternative possibility* open to her at the time of action. Specifically, this alternative possibility is such that it is not compatible with determinism.

The name given to this view derives from the fact that the agent must have some kind of ‘leeway’ with respect to their choice in some given situation: according to this view, if determinism were true, there would be no leeway to do otherwise in a given situation,

because there would be no ‘genuine’ alternative possibilities.⁸ Thus, the central feature of this account is adherence to the ‘principle of alternative possibilities’ (PAP) as a necessary condition for free will.

By contrast, ‘source incompatibilism’ has been gaining traction since the introduction of Frankfurt’s ingenious method for constructing counterexamples to PAP (so-called ‘Frankfurt-style Counterexamples’). This form of incompatibilism focuses much more closely on the causal origination of the action: it addresses the kind of ‘control’ the agent has over the action.

Frankfurt’s basic argument was that PAP is false: an agent can be morally responsible for an action even if she could not have done otherwise. I take no stand here on whether Frankfurt’s claim is true, although it has been widely accepted. Clearly, compatibilists have taken Frankfurt’s argument against PAP to support their position: if it turns out that alternative possibilities (at least the kind which are incompatible with determinism) are not in fact required for moral responsibility, then the leeway incompatibilist’s main argument in support of their position has been defeated, because they had claimed that an agent is free (‘morally responsible for their actions’) only if they have alternative possibilities in this sense.

⁸ Only, it seems, the truth of various counterfactuals.

However, source *incompatibilists* have also accepted Frankfurt's argument. They claim that what is important is not whether agents have alternative possibilities available to them, but whether they are the *source* of their actions. For example, they might accept that the imagined agent in a Frankfurt-scenario is morally responsible for her actions, even though (as Frankfurt points out) they could not have done otherwise, because that agent was still the *source of her action* in that scenario — and that's what matters for 'free will', not alternative possibilities.

The crucial incompatibilist move here is to insist that the relevant sense of 'ultimate source' is such that one cannot be the ultimate source of an action if determinism is true. According to Michael McKenna:

Source incompatibilists hold that determinism *does* rule out free will. But it does so, not because it rules out alternative possibilities, but instead, because, if true, the sources of an agent's actions do not originate *in* the agent but are traceable to factors outside her.⁹

⁹ McKenna (2003: 201-2). Emphasis in original.

McKenna suggests that, according to the Source Incompatibilist, determinism rules out free will because it would have the result that “the sources of an agent's actions do not originate *in* the agent but are traceable to factors outside her.”

The ‘origination’ that McKenna describes in this passage sometimes gets called ‘sourcehood’. It is this notion that I am challenging here. I claim that what is important about ‘sourcehood’ or ‘origination’ is really a feature of agency, but that it is misleading to think about these problems under the guise of such metaphors as ‘the source’ or ‘origin’. Hence, the terminology here obscures an important issue, and in my view the problem is better addressed by replacing such terms with that of ‘agential control’.

To briefly preempt the conclusion of this chapter, the point is that terms like ‘source’ and ‘origin’, and phrases like ‘originating *within* the agent’, are all suggestive of a certain special *location* within the agent. This is symptomatic of a decades-long fixation on causal determinism, where there is a felt need to ‘break the chain’ of causation that apparently leads ‘outside the agent’ into the deep past where no moral responsibility (or ‘free will’ for that matter)

existed.¹⁰ By contrast, my notion of agential control takes it to be causal *integration* rather than causal *source* that is important for agency.

To make things more concrete, I will now consider a particular example from the free will literature: Derk Pereboom's 'Four-Case Manipulation Argument' (4CA), which is billed as an argument against compatibilist views of free will, and which turns on questions of 'sourcehood' or 'origination' of this kind.

3.3.1 The Four-Case Manipulation Argument

The standard reference for Pereboom's 4CA is the version given in *Living Without Free Will* (2001).¹¹ Pereboom presents the 4CA as an argument against compatibilism: it is designed to show that determinism would rule out moral responsibility in a way that is

¹⁰ The point being, it seems, that where such chains are present, responsibility is transmitted along the chain from earlier times to later times. If there is no responsibility at earlier parts, because no human beings existed, then how can the agent be responsible at later times? So goes the general worry.

¹¹ The version found in Chapter 4 of Pereboom (2001) is the one I will refer to, but also see the more recent (2014). None of the changes in later modifications materially affect the following discussion here, and some of the commentary discussed below use the (2001) version, so I rely on that presentation of the argument.

similar to the covert manipulation of an agent. That is, Pereboom hopes to show that agents cannot be responsible for “decisions that are alien-deterministic events.”¹² He intends to show that

if an agent’s decision is an alien-deterministic event, he cannot be its *source* in the way required for moral responsibility.¹³

The reference to ‘*alien-deterministic*’ events here is important, because the problematic causal determination in these cases is determination by ‘sources outside the agent’: either in the case of covert manipulation by other agents, or, according to Pereboom, in a ‘normal deterministic world’.

In terms of structure, the 4CA is intended to be a general argument that works against any possible version of compatibilism, and in that sense it offers a ‘family’ of arguments, or a framework for constructing arguments in response to any new modifications to the compatibilist’s proposal. As Pereboom notes, the majority of compatibilist accounts are built on what he calls ‘causal integrationist conditions’ (CI-conditions). These conditions are

¹² Pereboom (2001: 90).

¹³ *Ibid.* My emphasis.

what I discussed in the previous chapter: they are what the compatibilist takes to be the conditions necessary for an agent to be morally responsible. Hence, having ‘free will’ according to the compatibilist is a matter of being an agent who meets the causal integrationist conditions.

Pereboom intends to show that compatibilist CI-conditions are not sufficient, because we can generate scenarios in which agents meet those conditions, but intuitively fail to be morally responsible. The upshot of Pereboom’s argument is that we must add a further condition—sourcehood—in order for the agent to be morally responsible, and this condition is not compatible with determinism. Hence no version of compatibilism can succeed, because the condition required for responsibility is specifically incompatible with determinism.

The 4CA claims to take any of the proposed causal integrationist conditions, and show that it is nonetheless possible for agents meeting those conditions to be *covertly manipulated*. Now while some might take this demonstration in itself to repudiate the compatibilist’s claim that those conditions are sufficient for moral responsibility, Pereboom goes on to construct a series of generalising cases which move from the case of covert

manipulation to the final case of a ‘normal compatibilist agent’ in a deterministic world. The point is to show not just that some particular version compatibilism is internally problematic (because the CI-conditions are consistent with manipulation), but that, if determinism is true, *all* our actions are like those under manipulation, because determinism is relevantly similar to covert manipulation.

It will be helpful to have Pereboom’s statement of Case 1 to work with, since this is the most important case, from which the other cases generalise. In the following extract I have highlighted the five CI-conditions that Pereboom includes, which he takes to be representative of the main compatibilist positions presently on offer in the philosophical literature:

Case 1. Professor Plum was created by neuroscientists, who can manipulate him directly through the use of radio-like technology, but he is as much like an ordinary human being as is possible, given this history. Suppose these neuroscientists “locally” manipulate him to undertake the process of reasoning by which his desires are brought about and modified – directly producing his every state from moment to

moment. The neuroscientists manipulate him by, among other things, pushing a series of buttons just before he begins to reason about his situation, thereby causing his reasoning process to be rationally egoistic. Plum is not constrained to act in the sense that **he does not act because of an irresistible desire** – the neuroscientists do not provide him with an irresistible desire – and he does not think and act **contrary to character** since he is often manipulated to be rationally egoistic. **His effective first-order desire to kill Ms. White conforms to his second-order desires.** Plum's reasoning process exemplifies the various components of **moderate reasons-responsiveness**. He is receptive to the relevant pattern of reasons, and his reasoning process would have resulted in different choices in some situations in which the egoistic reasons were otherwise. At the same time, he is not exclusively rationally egoistic since he will typically **regulate his behavior by moral reasons** when the egoistic reasons are relatively weak – weaker than they are in the current situation.

The 'generalisation strategy' employed in the argument begins with this case. Pereboom believes that everyone will have the intuition

that an agent in this scenario clearly would not be morally responsible for her actions, because of the covert manipulation. The argument then gradually moves through a series of slightly different cases until we get to Case 4, in which “physicalist determinism is true, and [the agent] is an ordinary human being”.¹⁴

The argument has been subject to a lot of discussion, and responses to it have generally been divided into two categories: ‘soft line’ and ‘hard line’ responses. The 4CA presents the compatibilist with a problem: if the agent (‘Professor Plum’) in Case 1 is clearly not responsible for his actions, and if Plum in Case 4 is just a normal agent in a world where determinism is true, then either Plum in Case 4 is not responsible (and compatibilism is false), or the compatibilist must point to some relevant *difference* between the cases, with respect to the ‘problematic’ determination. According to Pereboom, there is no relevant difference between the kinds of determination in each of the cases. Hence, the problem in Case 1 ultimately generalises to Case 4. The soft-liner has to show where, in the transition from Case 1 to Case 4, Plum starts being morally responsible.

¹⁴ Pereboom (2001: 115).

The soft line response to the 4CA thus shares the intuition that Pereboom is trying to evoke in Case 1 of the argument: the intuition that Plum is not responsible in that case, because of the manipulation. The soft-liner's task is to show that this case is importantly *different* from Case 4, which is just a case in which we suppose the truth of determinism, but where everything else is just as we normally suppose the world to be.

For example, the soft-liner could show that although Plum's responsibility is undermined in the early cases of manipulation, this is because Plum has *not* in fact met all of the compatibilist's CI-conditions. In other words, they must claim that Pereboom has not accurately described the compatibilist's vision of a morally responsible agent in setting up the 4CA. If these other conditions were added, this reply goes, then we would no longer have the intuition that Pereboom wishes to invoke, i.e. that such an agent is not responsible, and the 4CA could not get off the ground.

A *prima facie* problem with the soft line response is that the incompatibilist opponent simply adds in that new condition to their description of Plum in Case 1. The strength of the 4CA, and of manipulation arguments in general, is that they appear to be open-ended in this sense: whatever the causal conditions are that the

compatibilist deems necessary can simply be included, and a new form of manipulation devised such that the new conditions get ‘put into place’ by the relevant intervention.

For this reason, McKenna has been particularly influential in advocating a *hard line* response to the problem. In short, the hard line response turns the argument around, and aims to show that Plum in Case 1 *is* responsible for his actions, *because* the manipulation there is not relevantly different to determinism (as Pereboom argues) — and determinism is not a barrier to responsibility, by definition, for compatibilism. Here, the hard liner accepts the generalisation claim (that the determination in each case is not relevantly different), but takes this to show that the *responsibility* that is present in the deterministic case (which is just a statement of compatibilism) actually generalises to Case 1, featuring manipulation.

While this is an impressive attempt to bite the bullet, my argument will be that there is a deeper problem with the manipulation described in Case 1: it undermines Plum’s *agency*. Hence, as well as offering a more robust form of the soft line response, this will also serve to demonstrate the problem with thinking in terms of the ‘source’ or ‘origin’ of the action as being

important for agency, as Pereboom sets things up: instead, we shall see that these issues are about control, and a proper understanding of control reveals that it is *causal integration* rather than *causal source* that matters for agency. And while the notion of *sourcehood* is tied up with the standard ‘free will’ metaphysics of uncaused causes and metaphysical indeterminism, *causal integration* is straightforwardly approachable from the philosophy of mind, as I will explore further in later chapters.

3.3.2 Responding to the Four-Case Argument

While Pereboom obviously intends the 4CA to count against compatibilism, in the context of his overall position he intends the argument to count against event-causal *incompatibilism* as well. His own view is that only agent causal incompatibilism can account for free will and moral responsibility. However, according to Pereboom, agent causation is highly implausible given our best theories about the physical world, and hence we do not in fact have free will, and are not in fact morally responsible for anything (hard incompatibilism).

Nevertheless, it is Pereboom’s view that an agent causal incompatibilist account is the only view that *could* account for these

things, if it were empirically plausible. The principle on which this argument turns is what he calls a “claim about origination”, captured here by Principle O:

(O) If an agent is morally responsible for her deciding to perform an action, then the production of this decision must be something over which the agent has control, and an agent is not morally responsible for the decision if it is produced by a source over which she has no control.¹⁵

Pereboom argues that Principle O not only counts against compatibilism, as set out in the 4CA, but also that it counts against event-causal incompatibilism, since he believes that an agent could not have the requisite control over a *nondeterministic* event.

I will not consider Pereboom’s claims against event causal incompatibilism here, because it is the same argument that I discuss below in a more general context, when I consider the use of ‘agent causation’ as a means of responding to the various coherence objections. Pereboom, like other incompatibilists, does not believe that an agent could have the relevant control over a

¹⁵ Pereboom (2001: 4).

nondeterministic event unless they had a special agent causal power.

The important point to note here is Pereboom's suggestion that having the requisite control over the action requires that the agent be the causal *source* or *origin* of the action. If one thinks that what matters for agency is the causal source or origin, the theory of agent causation certainly seems to be a possible solution. However, I will argue that Pereboom's statement of the problem in the 4CA encourages a false dilemma, as a result of seeing the issue of control as one about origination.

By contrast, in the following chapters, I will develop an alternative account of agency. This metaphysics of agency fully expands on my notion of control: rather than seeing control in this sense as an *answer* to the 'problem of origination' discussed so far (as Pereboom encourages), it would be better thought of as a reframing of the problem itself, or a demonstration that a third alternative is possible, thus refusing to accept the dilemma set up by the 4CA.

Now, the aim in the present chapter is simply to show that the 4CA, which is typically presented as a problem for free will and moral responsibility in the light of determinism, actually rests on

certain assumptions about agency. To put this point the other way around: recognising that the metaphysics of agency is relevant here has implications for the philosophy of free will and moral responsibility,¹⁶ because when we have a correct theory of agential control, this automatically rules out certain kinds of manipulation, and invalidates a powerful argument in the free will literature — but not because of anything about moral responsibility, but because of our theory of *agency*.

If the manipulation in the 4CA suppresses agency, as I am suggesting, then, since any account of compatibilist CI-conditions is inseparable from a workable theory of agency, no compatibilist (or indeed *incompatibilist*) account could be subject to a manipulation argument of this kind. This is because agency itself is a matter of causal integration. The conditions that the compatibilist identifies as being necessary conditions for an agent to be morally responsible place various particular constraints on this causal integration, as Pereboom notes. But there is a more basic sense in which an agent must be causally integrated in order to be an agent at all: this is what the thesis of agential control says. Now there is an argument against the 4CA.

¹⁶ Of course, in my view, the addition of ‘free will’ here is redundant.

1. Agency is a prerequisite for anyone's account of moral responsibility or 'free will', compatibilist or incompatibilist, including Pereboom's own view.
2. According to the thesis of agential control, agency necessarily requires causal integration of a certain kind.
3. The manipulation described in Case 1 of the 4CA destroys this causal integration.
4. Therefore, the 4CA has failed to establish that there is a possible agent who satisfies the compatibilist's requirements on moral responsibility, *and* who is subject to covert manipulation, contrary to the stated premises of the argument.
5. The 4CA is invalid.

The task now is to show that the manipulation proposed in Case 1 actually suppresses agency. Someone who is subject to that manipulation would not be an agent: at the very least, they would not be an agent *with respect to* any behaviour occurring while they are subject to the manipulation, which, for the present argument, is the relevant occasion.

Kristin Demetriou has argued that the manipulation Pereboom describes in this case suppresses Plum's agency, and hence the argument fails as an argument against the compatibilist CI-conditions for free will and moral responsibility. Demetriou makes this claim in the service of a general soft line response to Pereboom.

In fact, John Fischer had already questioned whether Plum, manipulated as he is from his birth, could have ever developed a 'coherent self', and whether this is what undermines his responsibility (this not being the case in the straightforward deterministic world).¹⁷ The claim that Plum might not have a 'coherent self' is in some respects similar to the claim that he is not an agent. Clearly, then, other philosophers have recognised that there is something curious about the extent to which Plum is manipulated in the example, something that extends beyond 'merely' questioning his culpability.

I believe that the most important point to take from Demetriou's argument is that, while it may seem as though we can simply posit a kind of manipulation that brings about 'causal integrationist conditions' such as 'constancy of character' or

¹⁷ See for example, Fischer and Ravizza (1998: 234-5, n. 28).

‘having a reasons-responsive mechanism’, this is only because insufficient care has been taken to understand how these conditions relate to *agency*, which is itself a matter of causal integration. We can not simply assume that the proposed manipulation in Case 1 does not interfere with agency itself: as Demetriou argues, such manipulation is in fact antithetical to agency, and, for that reason, one cannot simultaneously suppose that such manipulation takes place, and that the compatibilist CI-conditions are satisfied. As she puts it (where PlumX is to be read as Plum-in-Case-X):

However, even if Plum1 and Plum4 have exactly similar physical and qualitative states, this does not ensure that Plum1 and Plum4 have the same status in terms of *agency*.¹⁸

To illustrate, she constructs a pair of diagrams, which will be familiar to those used to reading the literature on mental causation, to represent the differences between Plum1 (when manipulated by ‘neuroscientists’) and Plum4 (who is simply an agent in a deterministic world):

¹⁸ Demetriou (2010: 608).

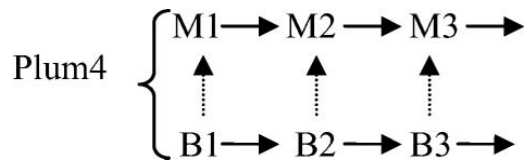


Fig. 3.1 Plum in Normal Deterministic World

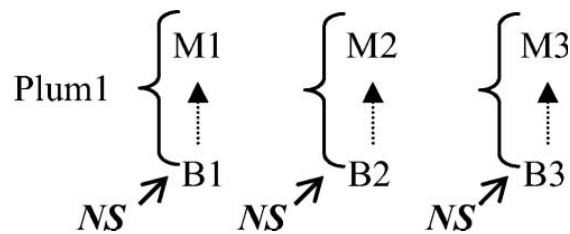


Fig 3.2 Plum Subject to Manipulation

In these diagrams, B represents the physical state of Plum (the “brain state”), M represents the mental state, and NS represents the neuroscientists who are manipulating Plum. The solid arrows represent causal relationships, and the vertical dashed arrows represent some kind of supervenience relationship.

The compatibilist holds that Plum4 is a normal agent in a deterministic world. While Demetriou does not go into great detail about this form of agency, enough is said to convey the general

features — the important thing is to compare it to the representation of what Plum’s agency would have to look like for the manipulation described in Case 1 to occur as intended.

Demetriou’s main problem with Plum1 is that he is not a causally integrated entity in the same way as Plum4, and so we ought to question whether ‘Plum1’ is really an agent at all.¹⁹ At least, there certainly appears to be something defective about Plum’s agency in this case. In Plum4, there is supervenience between the M level and the B level, but there is also causation across time: from B1 → B2 and from M1 → M2 and so forth. That is, *Plum’s earlier states causally impact his later states.*

By contrast, in Case 1, we are told by Pereboom that Plum’s states are brought about (caused) ‘moment-to-moment’ by the neuroscientists and their technology. The most charitable way of reading this claim for Pereboom’s argument, such that it does not turn the 4CA into a nonstarter, is that the manipulation will result in something like Plum1 above.

Demetriou goes to some length to consider the different ways of interpreting the ‘moment-to-moment’ claim, and settles on this as the most hermeneutically viable option. Briefly, if Plum’s

¹⁹ Demetriou (2010: 607).

own states *compete* with NS, then either Plum ‘wins’ (thus he is not manipulated, and 4CA cannot get started) or NS wins causal ‘control’²⁰ (and we have the diagram above). Similarly, if Plum and NS together *overdetermine* the causation of Plum’s states, then we do not clearly have manipulation either — since, by definition, Plum’s *own* state-causation would be independently sufficient for bringing about his later states.

Hence the argument can only get started if we assume that NS causes Plum’s states in the way represented in the diagram above. Of course, this diagram represents Plum’s own mental states here (M1, M2 . . .) as being *epiphenomenal*. Indeed, the diagrams in Figures 3.1-2 are similar to those texts in the philosophy of mind which address various difficulties for mental causation and epiphenomenalism.

The problem is that this kind of manipulation actually undermines agency. There is no causal integration between Plum’s own mental and physical states over time. There is a succession of

²⁰ Note that ‘control’ here means something like Pereboom’s understanding of ‘control’. One of my main arguments is that control is not a matter of standing in relation to *oneself* in something like the way NS stands to Plum.

‘time-slice’ mental and physical states, each individually caused by the external agents. As Demetriou puts it:

Plum1’s physical and qualitative mental states are not causally efficacious in bringing about his subsequent physical and mental states; Plum1’s states are, rather, the end effects of the causal powers expressed by the neuroscientists.²¹

The diagram for Plum4 tells us very little about what view of agency is being supposed here. One way of reading it is as a simplistic ‘causal theory’ of action, according to which actions are the result of behaviour being appropriately caused by intentional mental states. Demetriou may or may not have something like this in mind: the important point here is that *causal structure matters for agency*. Since the diagram for Plum4 is rather ‘bare-bones’, it is consistent with several more detailed views, including the one I will go on to present in later chapters. In principle, then, there is nothing wrong with Demetriou’s diagram for the agent in the deterministic world, it is simply under-described as it stands.

²¹ Demetriou (2010: 608).

The lesson to take from the 4CA is that the metaphysics of agency is more important to these problems than has previously been appreciated. In some sense, the Source Incompatibilist is right to question whether Plum could really ‘be the source’ of his action in these cases, since there is certainly something defective about Plum’s agency in Case 1, and it may be this which drives the intuition of non-responsibility here.

But the problem is not one of causal origination, it is one of causal *integration*. Pereboom’s argument would have us believe that we are genuinely responsible for our actions only if we can manage to stand in the same relation to *our* actions as the neuroscientists stand in relation to Plum’s. In Case 1, Pereboom supposes that everyone will share the intuition that Plum is not responsible, because the neuroscientists ‘are in control’.

Hence, in order to be genuinely responsible, Plum would have to occupy the position that is thus occupied by the neuroscientists: but this is impossible on the deterministic event causal picture, because there is always some prior set of states or

events which cause the later states.²² Therefore, the only way of satisfying this origination or sourcehood requirement (apparently needed for control), is by being a substance-cause of your actions for which there is *no* prior sufficient cause — hence nondeterministic agent causation, Pereboom’s favoured view of ‘free will’, emerges as the apparent solution to the problem he has set for the compatibilist.

By contrast, the view of agency presented in later chapters takes causal integration to be what matters for agency and control: in particular, that what matters is that there is the relevant causal integration between intentional mental states, consciousness, and behaviour. The point here is that focus on ‘origination’ is a red herring, brought about by a mistaken impression of what is needed for an agent to ‘control their actions’. As I will show in later chapters, causal integration of the right kind is sufficient to sustain a robust (and nonreductive) account of intentional agency, which is compelling in its own right.

²² This counts against deterministic and indeterministic event causal pictures, because in both cases there is always some prior event-cause, whether it causes its effects with probability 1, or < 1 .

Finally, before moving on to consider the second example from the free will literature, it is worth pointing out some implications of the above reading of the 4CA. In particular, there is an outstanding point regarding the ‘generalisation’ claim that is involved in Pereboom’s argument. So far I have argued that the 4CA cannot get started, because the manipulation involved in Case 1 suppresses agency, and hence the CI-conditions cannot be met, contrary to the stated assumptions of the argument. But the generalisation claim itself is an important part of the 4CA, and I have not directly addressed it.

The potential difficulty here is that one might accept that Pereboom’s proposed manipulation in Case 1 does suppress Plum’s agency, but also retain the generalisation claim: viz., that the manipulation in Case 1 is *not relevantly different* from determinism. Hence it appears as though the 4CA becomes an argument for agency incompatibilism! If the problematic determination in Case 1 is such that Plum is not an agent, and if the generalisation strategy remains valid, then all of our actions are like this if determinism is true, and agency is incompatible with determinism.

The response of course is to block the generalisation claim. As with the comments above, once we recognise that it is not causal

origination or sourcehood that matters, but rather causal integration, we can see that there *are* relevant differences between the kinds of determination involved in each case.

Note that the problem with Case 1 was not the fact that determinism is true: the problem is the kind of intervention on Plum's states carried out by the neuroscientists. Because of the manipulation, the causal integration required for Plum's agency is destroyed (all of his mental and physical states are caused independently of each other). Put simply, it doesn't matter whether they cause Plum's states deterministically or otherwise, the fact that they cause them at all (in this way) is what makes the trouble for agency.

By contrast, in Case 4 the 'determination' involved is simply the fact of physical determinism. There is no mention of external agents manipulating Plum. The exact causal relations that obtain will depend on the theory of agency that we are supposing, of course, but there is nothing about determinism which, in principle, interferes with causal integration, because deterministic causal integration is still integration.

3.5 Agent Causation

I noted above that Pereboom intends the Four Case argument to apply not just to the compatibilist, but also to the incompatibilist who takes an ‘event causal’ view of action. Although there is no single ‘event causal’ incompatibilist view, the general target here is any account in which the standard view of *action* is simply taken for granted. Since the standard view of action is generally supposed to be some kind of causal theory (the so-called ‘Causal Theory of Action’), such views are ‘event causal incompatibilist’ views. The ‘theory of action’ is the general view one has about how actions are produced, or come about: e.g. whether actions are events, whether they are caused by mental states such as belief and desire, and the relation between bodily motions and the action itself. Such questions are not specific to ‘the free will problem’, but are independent philosophical questions.

Hence the most basic way in which incompatibilists can disagree with compatibilists here is by insisting that the causation of behaviour by mental states, which for the compatibilist is sufficient for free and morally responsible action, must actually be an instance of *nondeterministic*—probabilistic—causation in order to be responsible action (i.e. ‘free’ as they put it). There need not be

any other disagreement other than regarding the presence of indeterminism (somewhere) in the story of action. This is perhaps one of the more common ways in which compatibilists and incompatibilists dispute the free will problem.²³ For example, Clarke writes, when comparing compatibilist views on ‘control’ with incompatibilist views:

My aim is to compare the degree of control that an agent might possess if her behaviour is deterministically produced with what might be possessed by an otherwise similar agent in the production of whose behaviour there is some indeterminism. The proper comparison requires that we look at agents who are *similar except for the indeterminism*.²⁴

The compatibilist standardly raises the coherence problem against such incompatibilist views. To see the force of this problem,

²³ This is ‘event causal incompatibilism’. See for example Robert Kane’s classic account in (1996). Wiggins (1973) gives an early and clear example. An interesting recent version of this type of view is Balaguer (2010). As for compatibilism, that view has almost never been associated with anything other than the standard view of action, and so is by default ‘event causal compatibilism’.

²⁴ Clarke (1995: 125). My emphasis, but Clarke’s sense of ‘control’.

consider van Inwagen's example of a thief who is torn between robbing the poor box, or refraining from doing so.²⁵ According to the incompatibilist view, it must be *metaphysically open* whether or not he will rob the box, right up until the point at which he acts. He has some reasons for taking the action, as well as reasons for refraining. There will be some probability attaching to each course of action, such that we can say he is *more likely* to do one rather than the other (or perhaps that they are equally likely).

Glossing over many of the details for the moment, we can see that when he acts, the desires and beliefs he had, which constituted the reasons he had for so acting, *caused* the action. Assume he chose to commit the crime. Then the reasons he had for doing so (a need for money, a belief that he could get away with it, etc.) are in fact the causes of the action. Had he refrained instead, the reasons he had for refraining would have been the cause of his behaviour. This is an instance of nondeterministic causation, but otherwise exactly matches the compatibilist's view of what happens when an agent acts. (Also note that it is not the theory of nondeterministic event causation per se that is the problem here — in general, such probabilistic causation is uncontroversial.)

²⁵ Van Inwagen (1983).

The point is that there are possible compatibilist and incompatibilist views of free will that *only* differ in respect of whether or not they require indeterminism. That is, only in terms of whether the casual connections between the relevant mental states and behaviour are deterministic or nondeterministic. They may otherwise be identical theories of *action*, including any epistemic or situational conditions that may be involved.

The coherence problem is this: since by hypothesis the incompatibilist agent may have reasons for doing A, and reasons for not doing A, it seems that once we have determined what the relevant probabilities are for each course of action, there is nothing else the agent can ‘contribute’ that further determines which action she takes, i.e. which of the various possible reasons *actually* cause the relevant behaviour. So it is in that sense ‘a matter of chance’ whether A or not-A occurs, and not ‘up to the agent’ which event occurs.

To put it another way, all of the ‘agential resources’ of the causal theory have been exhausted at the point at which the probabilities attaching to each set of reasons has been fixed.²⁶ By

²⁶ We need not assume that anyone can fix exact probabilities to the range of potential choices in actual practise.

hypothesis, the agent cannot have a *further reason* for settling which of the possible actions occurs, because her reasons have already ‘run out’ once their relative weightings have been established. There can literally be ‘no reason’ why the agent did A *rather than* not-A: that is, no way of answering the *contrastive question* why A rather than not-A. What *can* be given a reason is ‘why the agent did A’, namely, their reasons for doing A — and this would be true if they had done not-A instead, namely their reasons for refraining.²⁷

Hence, this type of challenge has sometimes been referred to as the problem of *reduced control*.²⁸ According to this view, the introduction of indeterminism into the causal story, compared to the otherwise-similar compatibilist story, can only *reduce* the control that the agent has over her action, because it introduces a (contrastive) explanatory demand that cannot be given a reasons-based answer — it cannot be given *any* explanation, other than ‘it was a matter of chance’.

In the case of determinism, it is suggested, this question cannot arise: the agent’s mental states cause the relevant behaviour

²⁷ See Levy (2005) for a forceful challenge to the incompatibilist along these lines.

²⁸ Here and while discussing the literature on ‘reduced control’, the sense of ‘control’ is not my specific sense of the term: it is used in the same way that the authors discussed here also use the term.

with a probability of 1, and there is no further *contrastive* question available. Therefore the compatibilist apparently does not face the problem of reduced control.²⁹ When the agent does action A for the reasons that they had, *that* is the answer to the question why the agent did A. There is no further contrastive question that can be asked.

However, this is not in fact the best formulation of the coherence objection, since it seems to commit the compatibilist to the truth of determinism. That is, it seems to make compatibilism *require* determinism. This combination of views has historically been called *soft determinism*, but most contemporary compatibilists take it as a virtue of their account that it is *compatible* with determinism, but does not require it.

Christopher Franklin has made this point explicit.³⁰ Rather than focusing on the potential *reduction* of control, he suggests that we focus on what he calls the ‘problem of enhanced control’ (although it would be better named the problem of ‘no

²⁹ The compatibilist might nonetheless say that the agent ‘could have done otherwise’ — see the footnote above (n. 23) in reference to the so-called ‘new dispositionalists’, e.g. Vihvelin (2013). Also see Clarke (2009) for the origin of that term.

³⁰ Franklin (2011b).

enhanced control’). The point here is that we may assume for the sake of argument that the introduction of indeterminism into the incompatibilist’s account does not reduce or eliminate control, since there are arguably ways of building indeterminism into the account which do not obviously threaten control: that is, they do not make the view obviously worse off than a relevantly similar compatibilist view.³¹

The important question is what indeterminism—considered by itself—could possibly add that *improves* the compatibilist account, i.e. that secures something important for free will (moral responsibility) that cannot be had by the rival compatibilist account. Since the *only* thing that the compatibilist cannot in principle help themselves to is indeterminism, it has to be shown how the bare fact of indeterminism turns a non-responsible agent into a responsible one. It does not seem as though the mere absence of something like deterministic causation can perform this kind of ‘alchemy’.³²

³¹ Although whether indeterminism in those places is *useful* at all is a separate question: that is the point of the problem of enhanced control. The ways in which indeterminism can be added which do not make the view worse off than determinism mostly seem to make the indeterminism irrelevant to the action.

³² Gary Watson uses this term.

At this point, the agent causal incompatibilist concedes that this does look like a problem for *event causal* versions of incompatibilism, but that their own view can accommodate it. They agree that, at best, the mere absence of determinism does not seem to secure any more control for the agent, and at worst it seems to threaten the coherence of the resulting view (reduced control).

However, the argument goes, agents that have an ‘agent causal power’ can be properly responsible for nondeterministic outcomes of this kind. Contrary to the challenge from the coherence problem, it is not simply the addition of ‘mere indeterminism’, but the introduction of the agent causal power *with* indeterminism that satisfies the incompatibilist’s demand.

Being an ‘agent cause’ in this sense is usually characterised by the notion of *substance causation*: on this view, agents *as substances* cause their actions.³³ It may be that the agent is the only cause of the action, or that, like Randolph Clark’s ‘hybrid’ incompatibilist view, the agent’s reasons are also partial causes of

³³ Actually, the more plausible versions of agent causation do not suggest that agents cause their *actions*, but rather that the agent’s action is *constituted* by the agent-as-substance causing some behaviour. I will sometimes speak of the agent *causing an action* for the sake of convenience.

the action.³⁴ In either case, the additional element is the involvement of a substance-cause that *is* the agent, and which *directly* causes the action qua substance.

Because of the special kind of ‘origination’ that being an agent cause confers, it is possible to be fully responsible for any nondeterministically caused actions, whichever one ultimately occurs. There is no ‘matter of chance’, since the contrastive explanatory demand is answered by pointing out that it literally was *the agent* who produced and originated the action — and this fact is supposed to undercut any further demands for explanation, including those which lead to the accusations of ‘chanciness’. To put it simply, when faced with the contrastive explanatory demand of why the agent did A *rather than* not-A, although a further *reason* cannot be given—by hypothesis—the fact that it was the agent qua *substance cause* that caused the action somehow *throws more weight behind that outcome*.

As I have suggested in response to Pereboom’s argument, the basic problem with the agent causal incompatibilist’s approach is that it fails to recognise that there is conceptual space between ‘the agent is in control’ and ‘the agent is the source / origin’. The

³⁴ Clarke (2003: Ch. 8).

assumption is that control just is being the source or origin of the action, i.e. in the way that Pereboom had supposed that the neuroscientists were in control of Plum. As we saw there, this naturally leads to the thought that the only way for the agent herself to be in control is to stand in that kind of relationship to her own actions — and that can apparently only be satisfied by being a substance cause of the action.

But this is also to misunderstand the nature of the ‘agent causal theory’, that is, what the real reasons are for accepting or repudiating the agent causal account. As Franklin has observed, the final step in the agent causal incompatibilist’s argument simply *assumes* that the agent causal power is only available to the incompatibilist.³⁵ But this is not argued for. Without this assumption, the response to the coherence objection fails on its own terms: the compatibilist could simply adopt the agent causal theory.

In that case, we would have an analogue of the original comparison between event causal compatibilism and event causal incompatibilism. The relevant comparison now would be between agent causal incompatibilism and *agent causal compatibilism*, where once again the *only* difference between the two views, in principle,

³⁵ Franklin (forthcoming).

is the bare fact of indeterminism. Thus the agent causal incompatibilist would face the problem of *no enhanced control*, because the agent causal compatibilist would also be able to throw metaphysical weight behind the causation of an action, beyond the ‘mere’ possession of reasons for that action.

In the original comparison between compatibilism and incompatibilism, the notion of ‘agent causation’ was invoked by the incompatibilist as a way of avoiding the coherence problem. But as Franklin points out, this only works if agent causation is only available to the incompatibilist, and whether this is so depends on the reasons for adopting the agent causal theory. If those reasons mean that the compatibilist can also adopt the agent causal theory, then the incompatibilist has lost an important dialectical move.

Franklin’s own answer is that in fact the best reason for considering agent causation of this kind is to avoid reductionism about the self. He argues that identity theories (e.g. a Humean ‘bundle theory’³⁶) and identification accounts (e.g. those offered by Velleman and Bratman³⁷) face certain internal difficulties, and the

³⁶ Olson (2007: Ch. 6).

³⁷ Velleman (2000); Bratman (2000).

agent causal theory is an alternative to those views, vis-a-vis the metaphysics of the self / agent.

Hence, according to Franklin, “if anyone should be an agent causalist, then everyone should be an agent causalist.”³⁸ That is, if the anti-reductionist arguments for agent causation are sound, then this is a reason for both compatibilists and incompatibilists to adopt the theory, because those reasons concern the reductionism, and do not specifically concern free will and moral responsibility.

My view is that the notion of ‘agent causation’ is a mistaken response to the problem of control — or better, it is an appropriate response to a mistaken conception of what the problem of control really is (origination vs. integration).

What is needed is to show how the *agent* is in control of her actions—simpliciter—and it is for this reason that the agent causalist gets something right in her criticism of the ‘causal theory of action’. Others have also observed that the so-called ‘standard story’ of action has a certain problem with accounting for the role of the *agent* in the production of action. Franklin’s reference to the reductionist theories of the agent given by Velleman and Bratman, therefore, is apt: there is a difficulty here, and the agent causal

³⁸ Franklin (forthcoming).

theory represents one possible response. Compare the identification response that is exemplified by Velleman and others:

The agent is moved to his action, not only by his original motive for it, but also by his desire to act on that original motive, because of its superior rational force. This latter contribution to the agent's behaviour is the contribution of an attitude that performs the functions definitive of agency; *it is, therefore, functionally speaking, the agent's contribution to the causal order.*³⁹

But the point here is that this problem—the problem of ‘self determination’, as Velleman puts it—is not simply a problem for ‘free’ agency, or ‘full blooded’ agency as Velleman (and indeed Franklin) have it: it is a problem for *agency as such*.

The problem of identifying the agent's contribution to the aetiology of action, indeed, how the *agent* fits in to the causal order at all, is not simply a problem that arises for some special instance of agency — whether that is the ‘freedom relevant’ part that occupies the literature on free will, or the ‘self determining’, full

³⁹ Velleman (2000: 141). My emphasis.

blooded’ exercises of autonomous, rational agency that Velleman and others are concerned with. It is quite simply the need for a theory of agential control.

With that in mind, we can see that the problem of ‘no enhanced control’ is not one that uniquely arises for the event causal incompatibilist. It is made *explicit* by the unique explanatory demands that are placed on that account, certainly, but it actually reflects a tension in the causal theory of action. In that sense, it is a problem for the compatibilist just as much as the incompatibilist, and this is where Franklin is exactly right.

Thus we get the strange ‘metaphysicalisation’ of Velleman’s functional role for the agent (viz. a *desire* to act in accordance with reasons, which thereby ‘throws weight behind’ the rational choice itself) which manifests itself as the agent causal incompatibilist’s response to the coherence problem that we encountered above — the thought that being a special substance-cause of one’s action somehow avoids the contrastive explanatory demand placed on the incompatibilist by giving the actual reasons-based outcome more ‘oomph’.

As for the problem of explaining agency vis-a-vis what is lacking on the ‘causal theory of action’, the agent causal theory is

one possible response to this problem, and has historically been invoked by incompatibilists about free will rather than those drawn to compatibilism. But as I said above, this fails to recognise the conceptual space between ‘the agent is in control’ and ‘the agent is the source / agent-cause’.

In the next three chapters, I develop an alternative theory of agency that shows how the agent can have control in a robust sense that is lacking on the basic causal theory of agency, but without invoking substance causation in the way historically associated with ‘agent causation’. It puts aside the mistaken view of control as ‘causal origin’, and focuses on what actually matters for agency, which is *causal integration*.

3.6 Conclusion

The previous chapter outlined my general argument against the existing dialectic of ‘the free will problem’. I claimed that there is no reason to retain the terminology of ‘free will’, since nearly all of the *substantive* issues that get discussed in those terms are shown to be about the conditions necessary for moral responsibility. ‘Free’ agency in this sense is just morally responsible agency, and ‘having free will’ is having a certain disposition or ability to act in a morally

responsible manner. Redundant though it is, however, the term ‘free will’ is best retained as a technical term, stipulated in the way that I have suggested above.

Rather than go on to develop a theory of moral responsibility, in this chapter I have picked up on the secondary issue that has played a role in the free will problem. This issue has been mistaken for an aspect of ‘free will’, one that does not apparently relate to moral responsibility in the way I just noted. In fact, it is about agency. Specifically, it is about a central feature of agency that I’m calling ‘agential control’.

I discussed two places in the free will literature where the notion of ‘sourcehood’ or ‘agent causation’ have played important dialectical roles, and suggested that we should instead see these issues as turning on the notion of control. In my sense, ‘control’ is something central for agency, and so both incompatibilists and compatibilists need to account for it. Hence, the incompatibilist has lost an important dialectical move against the compatibilist.

‘Source incompatibilists’ and ‘agent causal incompatibilists’ are mistaken to think that ‘being the source of an action’ and ‘being an agent cause’ are properties that are somehow unique to incompatibilism about free will: i.e., that they confer some special

advantage on the incompatibilist's view of free will. In fact, the only plausible way of reading these terms is such that they are ways of getting at the notion of control that I have been insisting on in this chapter.

In the next three chapters, I expand on the notion of agential control. In essence, the theory of agential control is really just the theory of agency itself, given how I have been using the term here. Control is central to agency. I said above that it is not 'causal origin' or 'source' (as Pereboom and others have claimed) which is important for agency, but rather *causal integration*. The theory of agency I develop in later chapters builds on this claim: agency is (roughly) a matter of the right causal integration between intentionality, phenomenal consciousness, and behaviour.

Chapter 4 makes a start at fixing the notion of agency that I am arguing for, and canvases the range of uses to which the term has been put. It picks up on Helen Steward's view of agency, which shares some important characteristics with the view I will go on to defend. It also offers an opportunity to consider the question whether agency is incompatible with determinism, which was briefly flagged earlier in this chapter. Then chapters 5 and 6 go on

to develop the two central strands of my theory of agency:
intentional realism, and phenomenal consciousness, respectively.

4

Folk Psychology and Agency

4.1 Introduction

I have already suggested that the theory of agency I am presenting has something important to do with ‘causal integration’. In fact, I went as far as to suggest that there is a sense in which agency *is* causal integration — of the right kind. Hence, one of the principal tasks ahead is to spell out exactly what ‘the right kind’ of causal integration is, and why that matters for agency.

First, I will make some introductory remarks on the notion of agency as it appears in the existing literature. In particular, one finds that the concept of agency is closely associated with two other concepts in the literature—that of *substance* (particularly substance causation), and that of *moral responsibility* (already noted)—and for that reason I wish to separate out, as much as is possible, the notion of agency itself from these two closely associated concepts. One way to do this is to look at the ‘limit cases’ in which agency is

strongly associated with one or the other of these related notions, and then to find an appropriate middle ground for talking about agency itself. Even if there are no actual philosophers who have taken quite these extreme positions on the matter, the point is that all of the features I mention are parts of actually existing theories.

First, there is the view according to which agency is strongly associated, perhaps constitutively connected, to the concept of *moral responsibility*. This view has historically been associated with the rationalist tradition, especially with Kantianism, and can be called the *moral agency view*.¹ According to the moral agency view, agents are the *bearers* or *subjects* of moral responsibility. Those things that are not possible subjects of moral responsibility (that are not ‘subject to the moral law’) are not, properly speaking, agents at all.

It should be clear that, given the foregoing discussion, this is not the view of agency that I have in mind. I do not believe that agency is constitutively connected to the notion of moral responsibility: this is far too *narrow* a use of the term ‘agency’ to capture much of importance beyond what is already contained in

¹ Although not put quite as bluntly as I have here, see Korsgaard (2009) for an example of this general line of thought.

the notion of moral responsibility itself. It seems to me that there is a very obvious sense in which it must be possible for something to be an agent without it being even a *possible* bearer of moral responsibility.²

As I said, perhaps no one actually endorses this view in the strictest sense, but it remains true that many accounts of agency seem to be weighted very heavily in the direction of morally significant agency, e.g. with a focus on the capacity to respond to specifically *moral* reasons. Part of my general argument here is to show that there is a sense in which agency can be usefully considered independently of these concerns, i.e. without bringing in the notion of morally significant agency, or moral reasons, *at all*.

Looking in other ‘direction’, as it were, we find the term ‘agency’ strongly associated with the notion of *substance causation*, i.e. the view according to which agency is primarily exemplified by simple cases of substance causation. We might call this the *substance view* of agency. According to the substance view, many inanimate substances are agents (in addition to people and other animals). For

² The question of whether it can be an *object* of moral significance is not the issue here. This is the distinction between a ‘moral agent’ and a ‘moral patient’, as it is sometimes put.

example, Alvarez and Hyman even go so far as to defend the view that such inanimate substances can be considered agents that *perform actions*.³ They characterise the term ‘agent’ very generally as something that ‘makes things happen’ causally. Hence, any kind of substance causation is at the same time ‘agent causation’, and any such substance is an ‘agent’.

On this view, agency is very widespread, because anything that has causal powers or liabilities is an agent (arguably, all substances). For example, a volume of acid is an agent when it dissolves a lump of zinc, and a bomb is an agent when it collapses a bridge. In fact, this view seems to make agent causation equivalent to substance causation.

Indeed, if one believes that there cannot be substances without causal powers or liabilities, then there is by definition no substance that is not at the same time an agent. On this view, of course, the notion of agency becomes rather empty, as it becomes impossible to distinguish between *agents* and things that are *not agents* (because there are none). I take it that if there is any commonsense understanding of agency, a large part of that

³ Alvarez and Hyman (1998).

conception is that not everything is an agent: some things are ‘just things’.

Consider the definition of agent causation provided by Lowe in a textbook on metaphysics, which is intended to capture this ‘substance’ view of the term:

An ‘agent’, in the sense intended here, is a persisting object (or ‘substance’) possessing various properties, including, most importantly, certain causal powers and liabilities.⁴

It is interesting to note that this ‘substance view’ of agency actually goes together, in a way, with the ‘moral agency’ view noted above. What I mean by this is that once the term ‘agency’ as been allowed to apply to the notion of substance causation in this way, the term becomes rather useless as far as making any important *distinctions* goes — yet such philosophers will no doubt still wish to make a distinction, as it were, between *us* and a volume of acid. And what might the important distinction consist in? The ability to respond to reasons, perhaps especially moral reasons. The volume of acid may

⁴ Lowe (2002: 198), and see in particular Chapter 11. See also Lowe (2008: Chs. 6-8).

be an agent (it may only be “an agent”), but we are *full blooded agents* in some important sense.

Where does this leave us? The point to take from this discussion is that thinking of agency in terms of substance causation as such, as some philosophers have done, leaves us without the ability to make important distinctions in places where there are strong practical and commonsense reasons to do so.

It thus turns out that thinking of agency along the lines of substance causation ends up leading back to the first way of characterising what is important about agency anyway: the focus on moral reasons, and specifically moral agency. My problem with that way of characterising things is that it is also too ‘narrow’. In fact, my full argument against the ‘moral agency view’ is found the whole of what follows, since I intend to show that there is a non-arbitrary way of characterising agency that does not depend on moral reasons.

In what follows, then, I will develop this non-arbitrary middle position: a useful concept that agency that stands on its own, such that we can understand and talk about agents that are not even possible moral subjects,⁵ while at the same time being able

⁵ Again, I do not suggest that such beings might not be *objects* of moral concern.

to clearly and easily distinguish between agency and ‘mere’ substance causation.

I would like to begin the discussion by focusing on Helen Steward’s view of agency. This is for two reasons. Firstly, the thesis of ‘agency incompatibilism’ was mentioned in the previous chapter as a possible consequence of the generalisation strategy employed by Pereboom. Now, although I argued against the generalisation strategy on its own terms, it is also worth saying a little more about why I believe that we do not currently have strong reasons to suppose that agency must be incompatible with determinism.

The second, and most important, reason for beginning with Steward’s view is that it is a highly interesting and plausible account of agency — incompatibilism notwithstanding. It falls neatly between the two extreme views that I just mentioned. It does not connect agency to the notion of moral responsibility in the way mentioned above, and neither does it limit the attribution of agency to the relatively small class of adult, rational human beings. On the other hand, it does not cast the net so widely as to leave us open to the problems noted above. So with that caveat, I believe Steward’s view of agency is an excellent place to begin.

4.2 Agency Incompatibilism

As part of her account of Agency Incompatibilism, Steward argues that we all share a certain innate concept of agency. As I understand it, the claim is that we are all in possession of this concept, and we ‘automatically’ apply this concept to certain features of our experience. Thus her argument at this point is not meant to establish a thesis *about agency*, but instead it is meant to establish a thesis about *our concept of agency*, i.e. basically a thesis about our psychology, and the way in which we apply certain concepts to experience. Steward claims that the agency concept has at least these four important features:

- (i) an agent can move the whole, or at least some parts, of something we are inclined to think of as *its* body;
- (ii) an agent is a centre of some form of subjectivity;
- (iii) an agent is something to which at least some rudimentary types of intentional state (e.g. trying, wanting, perceiving) may be properly attributed;
- (iv) an agent is a settler of matters concerning certain of the movements of its own body [...] i.e. the actions by means of which those movements are effected cannot be

regarded merely as the inevitable consequences of what has gone before.⁶

Steward's overall argument about agency is that: (1) the concept articulated above in points (i-iv) is part of our natural 'folk psychology', and that (2) if this concept of agency applies to anything at all, then it applies to many animals as well as human beings.⁷ As I see it, the 'agency incompatibilism' component of her theory of agency comes in primarily with point (iv).

I argue that Steward's defence of claim (1) is strongest with regard to points (i-iii), and for the claim that we do in fact have a concept of this kind, which may have evolved to enable us to categorise certain entities we meet with in experience differently to others. Indeed, I am also inclined to agree with claim (2), that we often apply something like this concept of certain nonhuman animals, but that we 'draw the line', conceptually speaking, roughly at this point — although at what specific point on the spectrum of living and non-living things we in fact draw the line is

⁶ Steward (2012: 71-2).

⁷ Steward (2012: 74).

probably not a question that we can answer.⁸ As part of (2), Steward makes a strong case for the claim that any good reasons for believing in non-reductive agency in the case of humans (usually called ‘agent causation’), also give us reason to believe in such agency in the case of other, *nonhuman* animals.

What has *not* yet been shown, I argue, is that the concept of agency includes a commitment to metaphysical indeterminism of the kind Steward claims is necessary to satisfy (iv). I argue that the evidence Steward provides is inconclusive for the claim that the concept of agency involves a commitment to metaphysical indeterminism.

4.2.1 The folk theory of mind and the concept of agency

In Chapter 4 of her book, Steward takes on the task of articulating what she believes the concept of agency involves. As already noted, she believes this includes at least (i-iv), although she does not rule out that there may be other features that she has not noticed.⁹

One early example involves imagining watching a large farm animal, such as a cow or sheep, engaged in its normal activities. She

⁸ I do not at this point take a stand on the question whether it is merely our concept, or the boundary itself, between agent and non-agent that is a ‘fuzzy’ one.

⁹ Steward (2012: 71, n. 4).

claims that a “normal and unprejudiced human being” will find it “almost impossible to avoid looking upon [the animal] as an agent.”¹⁰ For example, if the cow were to suddenly move from one side of the field to the other, we would find it quite natural to hypothesise that it did so because “the grass looked better over there, or because it was shadier, or because it wants to be nearer its calf, which has wandered off in that direction.” In other words, we would find it almost impossible to avoid interpreting that behaviour as intentional. We are also inclined to “think of such an animal as a creature that can, within limits, direct its own activities and that has certain choices about the details of those activities.”¹¹

Steward claims that our natural reaction to the activity of the animal discussed here is best explained by appeal to certain research in developmental psychology. This research claims that there are ‘domain-specific’ cognitive systems that are “designed from the outset to facilitate the application of mental concepts to certain of the entities we meet with in experience.”¹²

For example, it is claimed that these cognitive systems are implicated in answers to several important questions in

¹⁰ Steward (2012: 75).

¹¹ Steward (2012: 75).

¹² *Ibid.*

developmental psychology: e.g. the question of how young children can learn so much so fast, or how they can reliably generate the concepts necessary to learn a human language. Such domain-specific ‘modules’ are supposed to aid in the acquisition of such skills or knowledge. Steward argues that a similar processes must be at work here, playing an essential role in the attribution of mind and intentional behaviour to both fellow human beings and other nonhuman animals:

Both naturalistically inclined philosophers and developmental psychologists have argued that if infants and young children were restricted only to the sorts of reasoning and empirical evidence that philosophers have permitted themselves in attempting solutions to the so-called problem of ‘other minds’, it is impossible to see how they would ever manage to come by the system of interpretation by means of which the young child in fact effortlessly manages to encode certain motions as purposive actions, and treats them as revelatory of mental functioning.¹³

¹³ Steward (2012: 76).

This view of how we come to ‘encode certain motions as purposive actions’ is sometimes referred to as the ‘folk theory of mind’, ‘folk psychology’, or ‘intuitive psychology’. Steward argues that the concept of agency she defends is a central part of this folk psychology, and that it has a “modular basis in the human mind.”¹⁴

One of the main questions regarding this capacity to ‘mind-read’ is whether it should be regarded as involving a *theory*, analogous to the intuitive concepts deployed to explain the physical world (i.e. ‘folk physics’), or whether it should be seen as a kind of *mental simulation*, a form of modelling based on one’s own mental processes as an analog. Steward does not officially take a stand on this issue, but does note that the ‘theory theory’, as it is called, does not sit well with her other claims about the causal theory of action, because of the implicit tendency to view ‘mental states’ as unobservable particulars that are invoked to causally explain the occurrence of outwardly observable ‘raw behaviour’.¹⁵ Part of Steward’s claim is that it is not, properly speaking, certain ‘unobservable particulars’ that come to cause actions, but rather the *agent herself*.

¹⁴ Steward (2012: 71).

¹⁵ Steward (2012: 76-7; see also 77, n.13).

This is Steward's first substantive claim about agency. She claims that the conceptual framework that children acquire does not in fact view intentional activity in this way: i.e., as the causation of outward behaviour by mental states or events (as she believes the 'theory theory' is implicitly committed to).¹⁶

Although it is mentioned in several places, it is not clear how important the *modularity* claim is to Steward's account of agency. In any case I do not make any specific claim that the important concepts here must have a modular basis in the human mind. What seems to be the main point is that there is a concept of a *minded entity* which is the *subject* or *possessor* of certain mental states:

Such things as beliefs and desires, according to our folk theory, have to be *had*; beliefs require believers and desires desirers. This ownership relation, moreover, is not merely a matter of the states in question being located inside a given animal body; it is a matter of their being ascribed to something whose informational and motivational properties those states describe [...].¹⁷

¹⁶ It is not clear that the 'theory theory' is in fact committed to this—or any—view about the causal aetiology of action.

¹⁷ Steward (2012: 77).

She goes on to claim that the ‘minded entity’ which is the possessor of these mental states is *itself* regarded as the cause of the movements and changes in its body that constitute an action. This is opposed to the view according to which it is the mental states themselves (when they are conceived as particulars) which are the causes of those movements. This second point reflects a commitment to something like agent causation in the folk conception.

Finally, she adds the requirement for metaphysical indeterminism. This is part of feature (iv) indicated above, which requires that agents are *settlers* of various matters at the time of action. The term ‘settling’ is a technical term for Steward. An agent ‘settles a matter’ by acting: through performing the action, the agent *thereby* settles a matter that was not previously settled. The way in which this term is defined by Steward entails a commitment to metaphysical indeterminism of the kind that is not compatible with physical determinism.

To properly appreciate what is involved in this notion of ‘settling’, it will be necessary to consider the role that it is designed to play in Steward’s overall argument for Agency Incompatibilism.

After getting clear about the notion of settling, and the work it is supposed to do, I will return to consider the evidence Steward cites in defence of her concept of agency, especially for the defence of (iv).

4.2.2 ‘Settling’ and Agency Incompatibilism

The notion of ‘settling’ is central to the defence of Agency Incompatibilism. It is the feature of Steward’s concept of agency that she claims involves a commitment to metaphysical indeterminism. What is important about agency, according to Steward, is that certain things can be ‘up to’ an agent, in a way that nothing can be ‘up to’ a thing which is not an agent.

Considering an imagined reply on the part of the compatibilist, Steward denies that something being ‘up to someone’ can just be a matter of that thing depending causally on certain mental states of the person, such as choices or intentions:

Being an agent in respect of some particular action, that is, cannot simply be a matter of possessing certain internal states that bring about some relevant type of bodily movement ‘in the right kind of way’. For it is simply not necessary, in order

that some question or matter be up to me, that I should want a given outcome or intend it or choose it. What I have to be able to do, I shall argue, is to *settle* that matter.¹⁸

What gets settled is not a particular action or a particular event: it is the answer to a range of ‘questions’ that are *open* until the performance of the action. For example, Steward considers the action of buttering toast at a particular time. Strictly speaking, it is not *that action* which is ‘up to’ the agent, and neither is it *the action* that gets settled. It would not be right to say that ‘my buttering of the toast was up to me’. The right thing to say instead is that “the fact that there *was* such a buttering around the relevant time was up to me.”¹⁹ Hence, what gets ‘settled’ *by* the particular act of buttering the toast is the answer to questions that concern things like: whether or not there will be a toast buttering, when and where it will happen, whether I will use my right or left hand, etc.

The notion of something being ‘up to us’ is used by Steward to characterise what is special about agency. Agents are those entities that things can be up to, and nothing can be up to an entity

¹⁸ Steward (2012: 26).

¹⁹ Steward (2012: 37).

that is not an agent. In order to explicate this notion of ‘up to us’, Steward invokes the technical term ‘settling’. While no explicit definition of settling is given in the book, a useful reconstruction is given by Clarke:

(S) An action a that is performed at time t settles at t whether p iff (i) either it is impossible that a be performed then and the actual laws of nature hold and p , or it is impossible that a be performed then and the actual laws hold and not- p , and (ii) there is nothing existing at any time t' prior to t such that either it is impossible that that thing exist at t' and the actual laws hold and p , or it is impossible that that thing exist at t' and the actual laws hold and not- p .²⁰

In other words, the action a settles the question whether p if and only if the occurrence of a suffices for its being the case that p , or its being the case that not- p , and nothing else suffices for its being the case that p or not- p .²¹ It is thus important to note how the notion of ‘settling’ applies directly to the *world*, as it were, and not merely to

²⁰ Clarke (2014).

²¹ *Ibid.*

our knowledge of the world: we should be careful that the locution of ‘answering a question’ does not mislead in this regard. It is not that the question is ‘open’ because we *don’t know* if there was a toast buttering at some time *t*. Instead: *that there was* a toast buttering at time *t* is a fact that is settled by the performance of the action, along with the other relevant details of the action (whether it was done with the right hand, etc.).

On this point, we see why it is important for Steward that her notion of settling involves a commitment to metaphysical indeterminism. This is because Steward believes that, if determinism were true, for any given time there would be *some* prior conditions which sufficed for everything occurring at that time. As Steward puts it, everything would have been already ‘settled’ at the beginning of time (if there was a beginning). Therefore, if determinism is true, nothing can be settled by agents when they act, because those questions are already settled. Hence, nothing can be ‘up to’ an agent in her sense. Since it is constitutive of agency that *something* be up to the agent at some time (feature iv), this means that if determinism is true, there are no agents.

The first problem for this argument, as Steward acknowledges, is that it seems possible to understand the term

‘settling’ in a way that is compatible with the truth of determinism. And it is also possible to understand what it means for something to be ‘up to’ an agent in a way that is compatible with the truth of determinism. For example, one might point out that the falling of the third domino in a series *settles the question* of whether the fourth domino will fall, even though the falling of the first domino already ensured the falling of the fourth. One could explicate this point with the following counterfactual: if the third domino had not fallen, then the fourth domino would not have fallen.

The problem for Steward’s account is that it seems possible to find perfectly ordinary ways to use these words and phrases which do not entail any commitment to metaphysical indeterminism. Hence, merely appealing to such locutions as ‘up to us’, ‘settling the matter’, as well as other familiar terms from the free will literature such as ‘power to refrain’ and ‘open alternative’, will not by itself support the claim that agency requires the falsity of determinism, because both compatibilist and incompatibilist readings of these terms are available.

What Steward needs to do is *argue* that her stipulated way of understanding the term ‘settling’ is the right one for talking about agency. This is where the material in Chapter 4 of her book comes

in. As noted in the previous section, Steward aims to use the evidence from psychology, as well as appeals to intuition, to argue that we have a concept of agency which includes the commitment to metaphysical indeterminism. If this can be done, then she will be able to argue on that basis that her use of ‘settling’ is the right one for thinking about agency, because it is the one implied by our folk psychological commitments.

The empirical evidence regarding our core conception of agency will provide *evidence* in support of her claim. It will not, as she points out, constitute a “knock down argument”, but it will count in favour of her claims about agency. I will now argue that the evidence is at present inconclusive with regard to feature (iv), and thus it has not been established that our concept of agency contains this requirement.

4.2.3 Experimental evidence for an incompatibilist concept of agency

One of the most initially promising pieces of research that Steward considers is a paper by Shaun Nichols.²² In this paper, Nichols draws on experimental evidence to argue that young children in

²² Nichols (2004).

fact deploy a concept of ‘agent causation’. This concept has the following two features, according to Nichols’ account:

- (a) An agent is a causal factor in the production of an action.
- (b) For a given action of an agent, the agent could have *not* caused it. Roughly, the agent *could have done otherwise*.²³

The second claim is to be understood in a way that makes it incompatible with determinism. Hence, if Nichols’ argument is right, then Steward will be able to use this evidence to support her defence of feature (iv), because both involve a commitment to indeterminism. Unfortunately, as she notes, Nichols’ case for (b) suffers from the same problem that I argued affects Steward’s own account of incompatibilist ‘settling’: one cannot rule out compatibilist interpretations of the data that Nichols cites, and so it has not been established that the children studied do have an incompatibilist conception of ‘could have done otherwise’, rather than one that is compatible with determinism.

For example, one experiment reported by Nichols was designed to test whether children are committed to interpreting

²³ Nichols (2004: 475).

agents as possessing feature (b), while denying such attributions to inanimate objects. The point was to test whether children regard agents differently to ‘things’ in being able to ‘do otherwise’ than they actually do.

For instance, in one of the agent cases, children were shown a closed box with a sliding lid. The experimenter said, ‘See, the lid is closed and nothing can get in. I’m going to open the lid.’ At this point, the experimenter slid the lid open and touched the bottom of the box. Then the child was asked, ‘After the lid was open, did I have to touch the bottom, or could I have done something else instead?’ In the parallel thing-case, children were shown the closed box with a ball resting on the lid. The experimenter said, ‘See, the lid is closed and nothing can get in. I’m going to open the lid.’ At this point, the experimenter slid the lid open and the ball fell to the bottom. Then the child was asked, ‘After the lid was open, did the ball have to touch the bottom, or could it have done something else instead?’²⁴

²⁴ Nichols (2004: 483).

Nichols reports that every subject reported that the agent could have done otherwise, while all but one reported that the thing ‘had to’ do what it did. The immediate objection to interpreting these results as supporting incompatibilist features of agency is highlighted by Steward. She points out that no compatibilist is likely to deny that people view agents, when performing some action, as being able to do otherwise in a way that a *billiard ball* could not.²⁵ This result is consistent with everything a compatibilist might say is true about agency, so it is inconclusive.

Nichols is aware of this potential weakness, and thus goes on to discuss the findings from a second experimental set-up. This time the children were asked a similar set of questions about a human agent who has to make a choice between two flavours of ice-cream. The imagined agent is called ‘Joan’, and in the scenario described to the children she in fact chooses vanilla. The children are once again asked whether Joan could have done otherwise, or whether she had to choose vanilla.

However, this time the experimenters attempted to rule out compatibilist ways of understanding this question, by asking the children to imagine that “everything in the world was the same

²⁵ Steward (2012: 84).

right up until she chose vanilla,” and then asking them if she *had to* choose vanilla.²⁶ This case is then compared to two further cases in which the children are once again given the instruction to imagine that ‘everything in the world is the same’ until the event in question: first, one in which the example features a pot of water put on the stove to boil, and second, an event in which the agent performs a morally significant action (stealing a candy bar).

The results showed that more children gave deterministic answers for the physical cases than the moral ones. However, there was no significant difference between the responses in the physical cases and the ‘spontaneous’ one (choosing a flavour of ice-cream).²⁷

This second experiment does not support the claim that children deploy a concept of agency that is committed to metaphysical indeterminism, or which features a notion of ‘could have done otherwise’ that must be construed incompatibilistically. The first problem is the same one that I have raised already. It is not clear how the children understood the instruction to ‘hold everything in the world fixed’. This is a problem that Steward also raises against Nichols. For example, she questions whether the

²⁶ Nichols (2004: 487).

²⁷ Nichols (2004: 487-8).

children might have imagined everything was the same in the world, but not *in the agent*. Quite possibly, the children didn't realise that they were required to hold fixed the relative weightings of the agent's desires.²⁸

The second problem, of course, is simply that the results were inconclusive. Although there was a significant difference between the physical cases and moral cases, there was no significant difference between the children's responses to the agent making a choice about ice-cream and their responses to the water boiling on the stove.²⁹ This is especially disappointing for the proponent of incompatibilism about *agency*, because it is cases of such 'ordinary' choice that make up the vast majority of our actions as agents. *If* these results could be used to support any kind of incompatibilism, it would be some kind of incompatibilism about moral responsibility, but not agency.

Interestingly, two pilot studies conducted with adults, using the same experimental set-up, did reveal a statistically significant difference between these two cases. More participants claimed that the agent could have done otherwise when choosing ice-cream and

²⁸ Steward (2012: 84).

²⁹ Nichols (2004: 488).

that the water had to boil. However, as Nichols acknowledges, the high standard deviation in these results suggests that there is considerable variation among individuals with regard to their intuitions about these cases.³⁰

Although a little more promising, the difficulty here is that one would expect far less variation if these intuitions resulted from a ‘folk psychology’ that is supposed to be innate, roughly as Steward supposes.³¹ A wide variation such as this is better explained by the hypothesis that our concept of agency is *acquired* or learned, and is the result of various environmental, educational, or situational factors. Indeed, this would be much like the way in which Steward goes on to suggest that we acquire the ‘mechanistic world view’ that she claims has come to obscure our own *innate* conceptual framework (i.e. the innate folk psychology which characterises agency in the non-deterministic way that she proposes). She writes:

³⁰ Nichols (2004: 488, n. 7).

³¹ Nichols also raises a similar ‘nativist’ suggestion about the origins of this concept of agency, although he only considers it as one of several possible origin-explanations.

I regard the source of our deterministic intuitions as largely *cultural*. I believe we have them not because of any innate tendency to construe agency deterministically, but rather because of such things as the huge success of the seventeenth-century scientific revolution; the impressive results of sciences such as genetics and molecular biology, which have looked to explain properties of wholes in terms of properties of their parts; and the invention of the computer and the resulting temptations of mechanistic models of animal life.³²

Possibly, Steward could argue that the wide individual variation that Nichols observed with the adult study is in fact a *result* of the culturally-acquired deterministic intuitions obscuring our innate tendency to encode things as (indeterministically conceived) agents. That would explain why the higher variation was found with the adult group, but not with the children's group, given that the latter would not yet have had time to acquire the deterministic world view.

In any case, it is not enough to show that a concept of agency is *consistent* with the experimental data, as the indeterministic view

³² Steward (2012: 80).

of agency was already being considered as a possible ‘contender’. The point of citing this research was to swing the balance in favour of the incompatibilist’s view, by showing it to be the *more plausible* way to interpret the claims of ‘settling’, ‘up to us’, and ‘could have done otherwise’. I conclude that this has not been done and that the experimental evidence for the claim that folk psychology is committed to an incompatibilist agency concept is unconvincing at the present time.

4.2.4 Intuitions about the concept of agency

To recap the argument so far: It was objected that the term ‘settling’ (and a range of similar terms) can be used in ways that are not committed to indeterminism. Hence, it is not enough to simply establish the claim that agency must involve ‘the settling of certain matters’, or that genuine agency involves at least some matters being ‘up to us’. Assuming those claims are true, this nonetheless falls short of a defence of the specifically *incompatibilist* component of Steward’s claims about agency, because those claims can be perfectly well read along compatibilist lines.

The defence of the Agency Incompatibilism claim depends on being able to argue for the *empirical* claim that folk psychology is

committed to a concept of non-deterministic agency.³³ If true, this would support the claim that we should understand key terms like ‘settling’ in an incompatibilist sense, and would therefore support Agency Incompatibilism in general.

Other than the empirical evidence from psychology that was considered in the previous section, Steward also appeals to “the intuitions of individuals about what *their* folk psychology appears to involve [...] and the not inconsiderable evidence that is provided by the historical persistence and recalcitrance of the free will problem.”³⁴

Whether the historical persistence of the free will problem supports Steward’s view that our folk psychology involves a fundamental tension between determinism and agency in the way that she imagines is questionable. Indeed, one of my main arguments in this project is that the free will problem is in need of restructuring, not least because of its historical persistence. On the contrary, the history of the free will problem demonstrates that there is no clear consensus about whether determinism is a problem

³³ Steward (2012: 548-50). Here and elsewhere when I say ‘non-deterministic agency’, or something similar, I mean ‘a concept of agency that is such that agency is not compatible with determinism’.

³⁴ Steward (2012: 80).

for anything—recall Peter Strawson’s remark that he does not know what the thesis of determinism is³⁵—and that people’s intuitions about free will, moral responsibility, agency, determinism, etc., probably differ widely.

Nonetheless, some people do have incompatibilist intuitions, and Steward is one of them. She claims that some of these are expressed in Chapter 4 (before turning to discuss Nichols’ work). What if anything do they tell us about folk psychology, and how we should conceptualise the activity of agents? I will argue that, when it comes to the kind of determinism that is in question in debates about action, it is all too common for there to be a conflation (or ambiguity) between the notions of determinism and *reductionism*, and that this ambiguity itself may contribute to the very historical persistence of the free will problem that Steward notes.

Indeed, at one point, Steward seems to come close to claiming that it is not in fact *determinism*, but a certain mechanistic, or reductionist, view of human beings that is the source of the problem she is discussing. She writes:

³⁵ Strawson’s famous opening line from his now classic essay ‘Freedom and Resentment’ (1962: 1).

It is not universal determinism *per se* which really constitutes the thing with which agency is most specifically in tension. It is rather a more localised variant of the thesis, born of a conception of agency itself as a phenomenon which must be neatly superimposable over, or at least very straightforwardly supervenient upon, the various intuitively lower-level and impersonally describable phenomena that we know have something very important to do with it, which generates the real problem.³⁶

Of course, Steward remains an incompatibilist, and she does claim that determinism itself is a problem for agency. But there are several points at which she appeals to intuitions to support this claim, where the intuition being invoked is either more to do with reductionism, or is equivocal between the two (that of determinism, and that of reductionism).

Return to the main example given earlier in Steward's discussion of our folk concepts, that of a large farm animal moving across a field. Steward writes:

³⁶ Steward (2012: 10).

It is most unlikely that anyone not already encumbered by theoretical prejudices would suppose it had been caused to make its trek across the field by a strictly reflex action or a simple stimulus–response mechanism. The activity of a cow or a sheep, I suggest, simply does not *look* as though it could be explained by such means.³⁷

This appears to be the expression of an intuition (“simply does not *look* as though ...”). Yet from this point, the conclusion reached is that we must instead view the animal’s activity according to the indeterministic conception of agency.

I submit that it goes deeply against the grain to suppose that each exact detail of each movement orchestrated by an animal was settled at any point prior to a period broadly concurrent with what we think of as the period of the animal’s action.³⁸

However, the intuition expressed in the first extract is about certain forms of reductionism, or mechanism. That the animal’s activity is

³⁷ Steward (2012: 75). Emphasis in original.

³⁸ Steward (2012: 75).

not the result of “a simple stimulus-response mechanism,” or a “strictly reflex action.” Both of these points are neutral with regard to determinism (or indeterminism).

Steward cites the historical persistence of the free will problem as evidence that people do possess strong indeterminist intuitions of the kind she supposes result from possession of an incompatibilist concept of agency.

One way of bolstering the argument, though, is by appealing to the very persistence of the free will problem, and the ease with which it can be explained to the uninitiated, which might themselves be regarded as reasons for wondering whether any account of folk psychology that simply proposes that actions are events conceived of as being deterministically caused by prior mental states can really be quite right.³⁹

In this passage, it is not clear whether the intuition being expressed is:

³⁹ Steward (2012: 79).

- (A) Any account of folk psychology that simply proposes that actions are events conceived of as being **deterministically** caused [...] cannot be quite right.

Or:

- (B) Any account of folk psychology that simply proposes that actions are events conceived of as being [...] caused **by prior mental states** cannot be quite right.

So we seem to have a situation where, if one were to agree with the claim in the original text, this might indeed be because one has incompatibilist intuitions about actions being deterministically caused events — but it might be because one has anti-reductionist intuitions about actions being a matter of causation by prior mental states.

A similar ambiguity occurs here:

[M]ental states are simply not thought of by our folk psychology, I maintain, as independent causally efficacious entities. They are thought of rather as features of a *substantive* entity—an agent—which must *act* if any bodily movement is

to result from its desires and beliefs and whose actions are thought of as explicable by appeal to, but not as deterministically caused by, those desires and beliefs.⁴⁰

Compare:

- (A) The claim that mental states are not independently causally efficacious entities, but rather features of a substantive entity, “whose actions are thought of as explicable by appeal to, but not as [...] caused by, those desires and beliefs.”
- (B) The claim that actions of the substantive entity are thought of as not *deterministically* caused by those desires and beliefs.

According to (A), the intuition being expressed is that the actions of a substantive entity are not *caused by* desires and beliefs (i.e. by particular mental states). It might not matter, for example, whether determinism is true as long as the actions of this entity are *explicable* by reference to desires and beliefs, but not caused by them.

⁴⁰ Steward (2012: 77).

On the other hand, according to (B), the intuition being expressed is that the actions of this substantive entity cannot be *deterministically* caused by — whatever it might be that causes them. The intuition here is that it is the bare fact of determinism that causes problems for action.

Indeed, both of these claims (A and B) appear to be running themes in Steward's book, and it is clear that the most hermeneutically viable reading of the last passage I just cited is to read Steward as intending *both* claims. The problem even here is that, when it comes to intuitions, it is hard to know which aspect is really driving the feeling of intuitive agreement when the claims are blended together in this way.

In any case, it is clear that Steward, and many other philosophers working on the free will problem,⁴¹ do report incompatibilist intuitions about the relation between human action and determinism. I confess to having such intuitions myself. But it is not clear to what extent these intuitions occur outside of the cabal of professional philosophers. The 'experimental philosophy' studies

⁴¹ For example, Vargas (2013) considers the probability that philosophers have mixed intuitions about this issue, with some harbouring both incompatibilist and compatibilist intuitions. Nichols (2006) makes a similar claim about 'folk' intuitions.

that have probed that question are limited in number, and usually require philosophical interpretation. The difficulty is that the notion of *incompatibilism*, as philosophers use the term, is itself a highly ‘philosophical’ concept (i.e. it requires an amount of philosophical competence to even understand a sentence containing the term).

Hence, I submit that the evidence from intuition is once again inconclusive, and does not support the claim that folk psychology is committed to indeterminism in agency. I noted above that some experimental studies found that adults have widely differing intuitions about these matters. The present argument shows, firstly, that it is very easy to conflate claims about determinism and claims about reductionism. If this is true, then there is further reason to be cautious when citing the intuitions people have regarding determinism, because they may in fact be intuitions about reductionism or mechanism, and not determinism.

Secondly, it shows that the subtlety of the concepts involved (viz. incompatibilism, and specifically *metaphysical* indeterminism in action) means that there will always be difficulty in surveying the general population about such matters without first requiring a minimal amount of philosophical competence or explanation, and

hence undermining the claim that these are ‘prephilosophical’, folk intuitions about the matter in question.

4.3 The Evidentiary Role of Folk Psychology and Intuition

In the previous sections, I have considered a range of evidence for the claim that there is a particular concept of agency that we are in fact committed to, as a result of certain cognitive modules that are a natural part of our development — a ‘folk psychology’ that is more or less innate. Steward claims that, properly understood, this folk psychology includes a conception of *agency* itself that, among other things, views agency as something that is not compatible with determinism.

In the previous section I raised several broadly empirical concerns: I doubted that we do in fact have such a conception of agency, based on the fact that the evidence from experiment, as well as from more traditional appeals to intuition, is not compelling. But there is a more general, methodological question to consider: what is the relation between the structure of our folk-psychological concepts, on the one hand, and actual *metaphysics* on the other? In this case, of course, the question concerns the relation between the

folk psychological conception of agency just mentioned, and the metaphysics of agency itself.

4.3.1 Debunking

Why examine folk psychology when the philosophical task is to develop a *metaphysics* of agency? For Steward, there are at least two reasons. As explained in Section (4.2.2), it is intended to support her specifically incompatibilist reading of the technical term ‘settling’. The task is to use the empirical data to support the claim that folk psychology is committed to an incompatibilistically-construed conception of ‘settling’, i.e. that the folk conception of agency is such that we naturally interpret agents’ actions as involving a metaphysically indeterminate process.

The role that the empirical data is supposed to play here is this: when deciding between which is the more plausible interpretation of the concept, the incompatibilist reading or the otherwise equivalent compatibilist reading, the fact (if it is a fact) that folk psychology is committed a concept which has as part of its content the incompatibilistically-construed notion of ‘settling’

thereby *supports* that interpretation of settling when running the argument for Agency Incompatibilism.⁴²

Secondly, I take it that this claim about folk psychology is to be used as part of the more general philosophical argument, directly in favour of Agency Incompatibilism. The fact that we have such intuitions about agency, and that we in fact experience certain phenomena (actions) *as* indeterminate processes (as the Agency Incompatibilist claims they really are), thereby *counts as evidence for the metaphysical claim* that this is how agency in fact works. Hence, the more general appeal to empirical psychology and cognitive science is meant to play a role in a philosophical argument about some metaphysical feature of the world. Specifically, it is meant to provide evidentiary support in favour of a particular metaphysical hypothesis.

The relation between cognitive science and metaphysics is the subject of some debate. In this particular case, the argument moves from (i) the fact that we are disposed to judge that X, or that we have an experience *as of* X in certain situations—which is taken to be the result of certain ‘hardwired’ features of our cognitive system—to (ii) the metaphysical claim *that* X. (Or to the claim that X

⁴² See, in particular, the argument in Steward (2014: 547-50).

is more likely to be true than some rival hypothesis.) But the assumption that discovering empirical facts about our cognitive systems or, more generally, our psychology, could be used as evidence in support of metaphysical hypotheses about the state of the world is at least open to doubt. And indeed I will now raise some doubts about that assumption in more detail.

Here is an alternative perspective on the relation between cognitive science and metaphysics: debunking. There is a famous line of argument in normative ethics which takes the existence of (e.g.) cognitive scientific evidence of the kind noted in (i) to actually count as evidence *against* the existence of the phenomenon in question (X). In ethics, this argument usually takes the form of an *undermining* or *debunking argument*.

For example, it might be shown that certain moral judgments, which are ostensibly taken to be the products of reasoned deliberation about the facts, are directly influenced by basic affective systems responding to seemingly irrelevant features of the situation in which the judgment is made, such as whether the subject is seated at a clean or a filthy desk.⁴³ Hence, it is claimed that the moral judgment is *undermined*, because we do not in

⁴³ Schnall *et al.* (2008).

general believe that the aesthetic condition of the desk is a morally relevant factor that should influence the judgment. In addition, the subjects were of course *unaware* that the condition of the desk was influencing their judgment. Here we have a case where facts about our psychology and cognitive systems seem to undermine the judgments we make about a certain phenomenon.

A more general form of this argument is the *evolutionary debunking argument*: the claim is that certain intuitions about moral problems, or the existence of general moral traits or virtues, such as altruism, can be shown to exist as a result of evolutionarily more ancient systems. If these systems evolved as a result of the pressure, as Richard Joyce puts it, for our ancestors to “make more babies” then this fact is supposed to undermine the normative force of such intuitions or virtues.⁴⁴ Joyce writes about moral thinking in general:

It is naive to assume that these natural prosocial tendencies extend to non-cognitive feelings, behavioural dispositions, inclinations, aversions, and preferences, but not to *beliefs*. But acknowledging beliefs under the influence of natural selection raises epistemological concerns, for the faithful representation

⁴⁴ Joyce (2006: 222).

of reality is of only contingent instrumental value when reproductive success is the touchstone, forcing us to acknowledge that if in certain domains *false* beliefs will bring more offspring than that is the route natural selection will take every time. Moral thinking could very well be such a domain.⁴⁵

Even more specifically, Joshua Greene argues that *deontological* ethical judgments are driven by primarily emotional responses (and hence not by reasoning), while *consequentialist* ethical judgments are the product of different psychological processes that are more ‘cognitive’ and, for that reason, are more likely to be the result of ‘genuine’ moral reasoning.⁴⁶ For example, Greene suggests that there is a (debunking) evolutionary explanation for the observed disparity in consensus about two

⁴⁵ Joyce (2006: 222).

⁴⁶ Greene (2008). For a similar view about the status of moral intuitions, an important line of argument is found in Singer (2005) and his classic book on the topic (now revised 2011). Singer does not argue against deontological ethics in the way Greene does: his more specific argument is that we ought to carefully discern which intuitions are the product of ‘distorting’ evolutionary pressures, and which are the product of reason, since we cannot do without intuitions altogether.

similar problems in normative ethics: the *trolley problem* and the *footbridge problem*.

The trolley problem is a familiar device in normative ethics, in which a person is asked to make a moral judgment about what they believe they ought to do in the situation that is described.⁴⁷ In this thought experiment, the person being tested is asked whether they ought to flip a switch that will divert a runaway trolley away from five people who would otherwise be killed. However, by redirecting the runaway trolley in this way, it will be diverted on to a second track where it will end up killing a single person. The ethical dilemma is whether to save the lives of five innocent people, at the cost of killing one innocent person. Green reports that the philosophical and popular consensus is that it is morally acceptable to do so.⁴⁸

In a second very similar case, sometimes called the *footbridge case*⁴⁹, the consensus appears to go the other way. In this thought experiment, everything is the same apart from the fact that the

⁴⁷ The classic source is Foot (1967). See also Thomson (1976).

⁴⁸ Greene (2008) cites Fischer and Ravizza (1992) for the philosophical consensus, and cites a number of papers in which people have been tested experimentally about these problems for the popular consensus.

⁴⁹ Or in the less-P.C. times of the 1970s, the ‘fat man’ case. See Thomson (1967).

person in question cannot simply flip a switch to divert the trolley. Instead, they must push a stranger off a footbridge into the path of the trolley, thereby stopping the train and saving the lives of five people (but killing the hapless stranger). Greene reports that most people think it is *not* morally acceptable to do so.⁵⁰

Assuming that it is true, as it seems to be, that people consistently reach different judgments about these cases, the philosophical problem is to identify a morally relevant factor between the cases that explains the differing moral judgments. After all, there can be no genuinely moral difference without some other kind of (non-moral) difference between the cases.⁵¹

However, Greene suggests that there is a quite different explanation of this difference to be found. He argues that the different intuitions about these cases are due to the difference in the proximity of the violence that occurs. In the trolley case, everything happens at a distance and all the person has to do is flip a switch.

⁵⁰ Greene (2008: 43).

⁵¹ This is the notion of *moral supervenience* — usually attributed to R. M. Hare who was one of the first to use the term ‘supervenience’ in this context, although the idea itself has been around for a long time. See Moore: “one of the most important facts about qualitative difference...[is that] two things cannot differ in quality without differing in intrinsic nature” Moore (1922: 263).

However, the footbridge case involves physically pushing a stranger from a bridge and effectively killing them with one's own hands.

Given that the existence and threat of such 'personal violence' is evolutionarily old, it is plausible that we have strong negative emotional responses to such cases, owing to the necessity of living in close proximity to other creatures "who are capable of intentionally harming one another, but whose survival depends on cooperation and individual restraint."⁵² By contrast, the trolley case triggers no such responses, because the possibility of such scenarios is a relatively recent phenomenon. As such, Green suggests that people tend to deal with the trolley case in a 'more cognitive' way, and employ a form of cost-benefit moral *reasoning*.

Hence, the argument from cognitive science is this. Because it has been shown that the intuitions and the judgments that people make about this situation can be explained by reference to psychological systems that evolved to deal with problems of survival and reproduction, this undermines their status as reliable sources of information about the world (including the moral facts about the world, in this case). As has long been pointed out, it is

⁵² Greene (2008: 43).

difficult to see why evolution would have selected for systems that are capable of discerning objective moral truths.⁵³ To be precise, the debunking argument challenges the justificatory credentials of the *process* by which the intuitions or the beliefs in question are formed.⁵⁴

In fact, the last point made above is more properly directed at arguments in meta-ethics, since it concerns the objectivity of moral beliefs: the general form of argument is that there is no reason why evolution would produce systems that are reliable trackers of objective truth, especially objective moral truths, rather than being sufficient only for the more limited instrumental goal of survival. As Schaffer puts it:

Evolution suggests that human cognition is a powerful but flawed tool. On the one hand it is plausible that many of our cognitive faculties evolved to help us with the four ‘f’s (feeding, fighting, fleeing, and reproduction), and plausible that this pressured our ancestors towards reliably tracking the environment. On the other hand it is equally plausible that

⁵³ For an important argument along these lines, see Street (2006).

⁵⁴ See Nichols (2014) for the distinction between ‘best explanation debunking’ and ‘process debunking’ of this kind.

many of our cognitive faculties evolved to give us quick and dirty heuristics reliable only for limited purposes in evolutionarily salient contexts.⁵⁵

The undermining arguments just outlined suggest that the evidence from cognitive science that Steward appeals to could actually be taken as counting against the claim that agency works in the way our folk psychology appears to suggest. The experiences as of agency, or the intuitions we have about agency, are driven by basic systems which have evolved to track various features of the environment, and categorise them differently to others. But do they reliably track real, metaphysical features of the world? Or are they instead the result of ‘quick and dirty heuristics’ that are ‘reliable only for limited purposes in evolutionarily salient contexts’?

Is Steward open to a debunking argument? In addition, even if the claims about agency are not suitable for *debunking*, what exactly are the epistemic credentials of such sources of information in general? That is, to what extent can we rely on these experiences and intuitions (the deliverance of our folk psychology in general) as sources of evidence in support of a metaphysical hypothesis? What

⁵⁵ Schaffer (2016: 342).

weight should we assign them as sources of evidence, if we allow them to count at all?

4.3.2 The relation between cognitive science and metaphysics

In general, information of this kind will be relevant to metaphysical inquiry. However, the kind of *unselective* debunking approach noted above is not appropriate. Even in normative ethics, among those who have criticised reliance on moral intuition, such as Peter Singer, the claims are appropriately selective: Singer argues that we cannot do without intuitions altogether, and so it is misguided to simply dismiss all reliance on intuitions as sources of information. Instead, what we must do is attempt to discern which intuitions or experiences are the product of ‘distorting’ evolutionary pressures, and which have a rational basis.⁵⁶

This selective approach is the right way to assess the relation between cognitive science (and psychology, neuroscience, etc.) and

⁵⁶ Singer (2005). He writes that we must “attempt the ambitious task of separating those moral judgments that we owe to our evolutionary and cultural history, from those that have a rational basis. This is a large and difficult task. Even to specify in what sense a moral judgment can have a rational basis is not easy. Nevertheless, it seems to me worth attempting, for it is the only way to avoid moral skepticism.” (2005: 351)

the evaluation of metaphysical claims about the world. For example, the *mere fact* that some feature of our experience (e.g. we experience certain entities as *agents*) can be shown to come about as the direct result of a cognitive system that has been shaped over many years of evolution, under the pressures of reproduction and survival, does not for that reason serve to debunk or undermine that feature of our experience (or that intuition, or that belief ...). In fact, *all of our experiences and intuitions are enabled by cognitive systems which are the product of such evolutionary pressures* — because human beings are the products of evolution. Again, this is a point that is put succinctly by Schaffer, in answer to the question ‘when to debunk?’

One *bad* answer—bad because unselective—is that an intuition can be debunked when one can tell a cognitive story about how it arises. This is unselective (and thus bad) because of course there is always some cognitive story to be told about every cognitive output, intuitions included. Cognitive outputs are not miracles. They all have causal aetiologies through our cognitive engines.⁵⁷

⁵⁷ Schaffer (2016: 344).

Secondly, and this is a point that Steward makes herself, the mere fact that a system can be triggered in error does not thereby undermine the justificatory force of that system in general (which would be a kind of *process debunking*). We can and should regard the outputs of such systems as being defeasible. But if the system is triggered in error all the time, more often than it is accurate, then we would likely regard it as a unreliable process of belief formation in general, and this would support a debunking argument. In normal circumstances, however, we accept that the possibility of error does not automatically undermine a particular source of information about the world: so, for example, the fact that we are subject to the possibility of perceptual errors or illusions does not undermine perception as a source of knowledge *in general*.

What is the right way to treat the evidentiary credentials of our experiences? In particular, how should we evaluate the agency case, where the claim is that we have an experience of other agents (and, in particular, where we experience these agents as instantiating the various properties that Steward has outlined)?

Firstly, the fact that the modular systems that Steward suggests are responsible for these experiences (and perhaps also for

the intuitions she cites) are products of evolution does not by itself present a problem. It would need to be shown that either (i) the *best explanation* for our having these experiences in the way that we do can be given in terms that do not presuppose the truth of the agency theory (as we saw in the case of the filthy desk affecting the moral judgments, for example).

But this has not been done, although I do not rule out the possibility that such challenges could be developed. This would be a further empirical matter.⁵⁸ (The fact that the concept of agency in question here is one that applies almost exclusively to other *animals* is cause to speculate whether this system evolved as a way of categorising those features of the environment that are potential threats.) Hence, I conclude that there is no ‘best explanation’ reason to dismiss these agency experiences.

Secondly, the way in which we come to have these experiences is grounded in the perception of various features of the environment. That is, the experience of agency Steward describes appears to be fundamentally a *perceptual* one. Although this is not entirely clear from the text, I believe it is a plausible reading of the

⁵⁸ Again, see Nichols (2014) for an explanation of the difference between ‘best explanation’ debunking (or undercutting / undermining) and ‘process’ debunking.

account: see, for example, the discussion above about Steward's thoughts on the 'theory-theory' vs. the simulation approach (which she favours), and her comments on folk psychology as a 'way of seeing':

"This way of thinking is, moreover, at the same time a way of *seeing*"⁵⁹

"As I suggested above, one normally *sees* as cow as an agent. One does not *judge* that on balance, beliefs and desires present the best explanation of its behaviour."⁶⁰

For this reason it is not plausible to suggest that the *process* by which the beliefs are formed is an illegitimate one. For example, Goldman suggests that there are some kinds of process which are 'essentially' faulty, and not legitimate processes of belief formation: "confused reasoning, wishful thinking, reliance on emotional

⁵⁹ Steward (2012: 93). See also the reference to Wittgenstein: "We say 'The cock calls the hens by crowing' ... Isn't the aspect quite altered if we imagine the crowing to set the hens in motion by some kind of physical causation?" (1953: §493; cited in Steward 2012: 93 n. 37).

⁶⁰ Steward (2012: 77 n. 13).

attachment, mere hunch or guesswork, and hasty generalisation.”

By contrast, forming beliefs on the basis of direct perceptual experiences is not in general a faulty *process* of belief formation.

One might claim that the perception of agency is in fact a perceptual illusion, thereby attempting to block the move from perceptual experience to the justification of beliefs. However, this would require a specific independent argument for the claim that agency perception is in fact an illusion. As we have seen above, the mere fact that perception in general is open to the possibility of perceptual illusions does not undermine it as a source of knowledge. That we might be subject to an illusion of agency in this or that case would not undermine the general claims about agency perception that Steward makes. It would need to be shown that we are *systematically* subject to a perceptual illusion in the case of agency.

Perhaps a better way to challenge the claim about agency perception would not be to dispute the *process* that leads to the belief, but to deny that the experience in question actually has the content that Steward claims it has. However, this would not be a debunking or undercutting argument any longer: it would be a straightforward disagreement about the content of that experience.

Indeed, I suggested above that such experiences do not plausibly have metaphysical indeterminism as part of their content: but my argument there was simply that there is insufficient empirical evidence to suggest that most people have an experience of that kind, on the basis of their introspective self-reports, and the availability of research which tests people's intuitions about various problem cases.

A deeper philosophical point to make in this context, when evaluating the evidential status of our experiences, is that we should be aware of the possibility that introspective reports of experiences may be theory-laden, especially when the reports issue from professional philosophers working on (for example) the free will problem.⁶¹ More importantly, Horgan and Timmons argue that even if the phenomenology genuinely has aspects that are aptly described in incompatibilist terms—as we saw with Steward's feature (iv) which, it was claimed, involved metaphysically indeterminate settling of matters at the time of action—this does not by itself mean that the phenomenology thus characterised has incompatibilist *satisfaction conditions*. They write:

⁶¹ Nahmias *et al.* (2004: 163).

[I]t is one thing for the phenomenology to be aptly described as ‘an experience as-of having an unconditional ability to do otherwise in my actual circumstances’; it is another thing for phenomenology that is thus aptly described to have libertarian satisfaction conditions involving the falsity of determinism. Problematic theory-ladenness might intrude itself not in the use of the phenomenological descriptions themselves, and not in the aptness of these descriptions in characterising the phenomenology of freedom, but rather in one’s construal of the intentional content of the pertinent phenomenology as thus described — its satisfaction conditions.⁶²

Hence, even if it were granted, with Steward, that we are indeed naturally disposed to have experiences which are aptly described as experiences (as) of metaphysically indeterminate settlements at the time of action (and hence incompatible with determinism), this would not by itself entail that these experiences have incompatibilist satisfaction conditions — in short, it would not

⁶² Horgan and Timmons (2011: 188). They note that Nahmias *et al.* do not appear to recognise this second, deeper form of theory-ladenness.

entail that agency must in fact be non-deterministic in order for us to have the type of experiences that Steward describes (if we indeed have such experiences).

I argued above that there is not enough experimental evidence to suggest that the majority of people have experiences of, or intuitions about, metaphysical indeterminism being an essential part of agency. This is enough to dispute the claim that folk psychology is committed to an incompatibilist conception of agency, because the argument *for* that claim was based on empirical research and intuitions (dealt with in the previous sections). But the argument by Horgan and Timmons is an important one, and presents a challenge to the general project of citing experiences as evidence in the construction of metaphysical theories.

In particular, the contentious point in Steward's description of our agency experiences is the feature which is supposed to involve metaphysical indeterminism. Now suppose that this were in fact part of folk psychology. Considering that it is opposed to the rival interpretation on which agency is compatible with determinism, the contentious question here is whether the content of the agency-experience should be regarded as involving (metaphysical) indeterminacy which presupposes the falsity of

determinism, or a kind of indeterminacy that is compatible with metaphysical determinism.

It is difficult to see how we could register the difference between an experience as-of one or the other kind, without assuming that the experience is theory-laden in the way Horgan and Timmons suggest. Simply put, an experience as-of the action being indeterminate in a metaphysically robust way (settling a matter not previously settled) would *look the same* as an experience of that action being indeterminate in a way that is compatible with determinism (e.g. because I am utterly unable to predict the outcome, it is experienced as indeterminate). Along these lines, consider the story related by G. E. M. Anscombe about an encounter she had with Wittgenstein:

He once greeted me with the question: ‘Why do people say that it was natural to think that the sun went round the earth rather than that the earth turned on its axis? I replied: ‘I suppose, because it looked as if the sun went round the earth.’ ‘Well,’ he asked, ‘what would it have looked like if it had *looked* as if the earth turned on its axis?’⁶³

⁶³ Anscombe (1959: 151).

4.4 Conclusion

Here is what I believe we can say about the concept of agency. We have good reason to think of agency as something that is instantiated by *living* creatures (but not non-living entities), primarily exemplified by human beings — especially given the further features of agency that I will go on to consider in the next two chapters.

With that said, one of the interesting features of Steward's view, that I have not discussed here, is her claim that there is no *principled* reason why agency should be coextensive with human beings. Indeed, according to the present view, the extent to which the agency concept applies to nonhuman animals, for example, is going to be a matter of determining the extent to which they meet the criteria of the following chapters. Hence, the gradualism about agency to which Steward subscribes is itself certainly plausible, but it is not something that there is space to pursue in the present project.

For that reason, I will only consider the paradigm case, as it were, of agency: human beings. This should not be thought of as an anthropomorphic prejudice of mine — it remains an open question,

in my view, to what extent, and to what degree, agency is instantiated in nonhuman animals.⁶⁴ I consider human beings because intentionality and consciousness are probably best understood in relation to our own case, and it is enough of a challenge to argue that *human* agency is constituted by the integration of these phenomena, as I am doing here, without raising any further contentious questions. To that end, the next two chapters are devoted to unpacking each of these important aspects of the overall causal integration that constitutes agency.

⁶⁴ The question of degree, of course, applies also to human beings.

Intentional Realism

5.1. Introduction

In the previous chapter I made two important claims. First, that there is a ‘folk psychology’ which characterises our default way of interpreting the behaviour of other people, and often, to some degree or other, the behaviour of certain non-human animals as well. Secondly, that this folk psychology is relatively robust, although it is open to correction by empirical science in certain matters of *detail* — but not to the extent of a wholesale replacement of folk psychology by non-intentional theories of behaviour (e.g. at the level of neuroscience, for example).

In this context I considered Steward’s work on agency. Her work is relevant here because she makes a similar claim about the innateness of folk psychology as a means of understanding and predicting agents’ behaviour, and also, importantly, her view meets the general constraints on a concept of agency that I suggested at

the beginning of Chapter 4. Hence, it is useful to begin exploring the details of my view of agency by situating this discussion in the context of existing literature in the area.

A large part of Steward's account is focused on the relation between determinism (or indeterminism) and agency: she claims that indeterminism is a central part of our folk psychology — i.e. it is part of our innate conception of agency that it involves indeterminism. In the previous chapter, I expressed doubts that our folk conception of agency does in fact contain a commitment to indeterminism in this way.

A further set of claims that I discussed in that chapter concerned the general relation between such empirical facts about our psychology (i.e. how we in fact categorise certain things we meet with in experience) and straightforward metaphysical claims about the world. I suggested that we should take the existence of these cognitive dispositions (e.g. to categorise certain things as *agents*) to give us a—defeasible—reason for thinking that there is a real metaphysical distinction to be found here, between agents and non-agents, which maps the distinction made by our folk concepts — with the above-mentioned proviso regarding correction of those folk concepts *in matters of detail* by empirical science.

Accordingly, this chapter begins the positive task of articulating that view of agency in more detail. I will unpack the general view of agency as I conceive it, and will point out some similarities—and the important differences—between this view and Steward’s view of agency discussed in the previous chapter.

Aside from the doubts about indeterminism, I agree with Steward’s claims about the importance of attributing intentional states such as belief and desire as a central part of the attribution of agency. Hence this chapter addresses the question of intentional realism and how it forms part of our concept of agency. I will also consider two objections to the very idea of intentional realism itself, i.e. those of *eliminativism* and *instrumentalism*.

While the problem of eliminativism is relatively quickly dispatched, the challenge presented by instrumentalism leads into a deeper discussion of causation and its role in our concept of agency: it is not simply intentional realism that matters, of course, but intentional *agency*. The conclusion of this section is that the interventionist view of causation provides a way of responding to the instrumentalist — but in a way that must involve a causal contribution from phenomenal consciousness. Thus we come full circle, back to the notion of agency as involving both intentional

realism and consciousness, in a way that cannot be reduced to the individual contribution of either. The following chapter then takes up a deeper discussion of consciousness and its role in agency.

5.2 The Agency Concept

Explanation of action by reference to intentional states is central to our conception of agency, and those intentional states are characterised by the structure of our folk psychology. Thinking of something that can *believe* and *desire*, and can take action to achieve a goal or outcome that is desired, is to think of that entity as an *intentional agent*. For example, Kim famously sums up the continuing interest in the problem of mental causation by pointing out that “we care about mental causation because we care about agency.”¹ Elsewhere he goes on to say:

Let us first review the reasons for wanting to save mental causation—why it is important to us that mental causation is real. First and foremost, the possibility of human agency, and

¹ Kim (2010: 257).

hence our moral practise, evidently requires that our mental states have causal effects in the physical world.²

Those mental states that Kim takes to be centrally involved in human agency are those such as belief, desire, intention, and so forth — those which make up folk psychology. Kim also notes that it is important for those mental states to have *causal* power. In fact, Kim makes it clear that in order to be a realist about mental states—or indeed about anything at all—one must be committed to the *causal* reality of those states. This he calls ‘Alexander’s Dictum’: to be real is to have causal powers.³

Even those who do not hold a causal view of action believe that intentional explanations are nonetheless centrally involved in action explanation. While those non-causal views face a number of difficulties that are familiar in the literature, particularly since Davidson’s seminal 1963 paper, I do not propose to rehearse them in any detail here.⁴ The point is that they nonetheless depend on

² Kim (2005: 9); see also (1998: 31) for similar remarks.

³ Kim (1993: 348).

⁴ Davidson (1963). A prominent recent defence of a non-causal view of action is Ginet (2007). See Clarke (2010) for a nice overview of the criticisms of Ginet’s non-causal view, and why his responses to them do not succeed.

such intentional explanations as a central part of how we understand human action.

Notwithstanding the many differences of detail between the different accounts, the received view is currently that reasons are (in some sense) causes: that is, the intentional explanations we give, involving the folk psychological predicates under discussion here, are *causally* involved in the explanation of action. The general understanding is that—as Davidson pointed out—if reasons were *not* causes, we could not explain why an agent performed a particular action A for reason *p* rather than reason *q*, where both *p* and *q* are reasons to A.

If, as the non-causal view would have it, we have an intentional explanation on the one hand, and a causal explanation on the other hand, then the reasons explanation could explain *why* the action happened (normative explanation), but not why the action happened *then* (causal explanation); likewise, the causal explanation could explain why it happened *then*, but not *why* it happened. This much is already familiar from the literature, and I suggest that we should agree with the principle that intentional explanations ought to be considered properly *causal* explanations. In any case, we can simply note that all parties to this debate agree,

at the very least, that intentional properties are somehow central to the explanation of action, and hence to *agency* — whether they are causally involved or not.

We therefore arrive at the first point of contact between agency and intentional realism: intentional explanations are central to the explanation of action. Recall from the discussion of Steward's work in the previous chapter that giving intentional explanations of behaviour is something that we do in certain cases (viz. most people that we encounter, and perhaps some other 'higher' non-human animals), but not in others. Hence, our folk concepts make a *distinction* between certain kinds of entity: namely, those which are properly categorised intentionally—those which are the subjects of intentionally-explicable behaviour—and those which are not.

If we are to use this *prima facie* distinction to support a metaphysical distinction, that is, to support a thesis about agency which in fact draws a distinction in reality between agents and non-agents, then we are going to require an account of what it takes to have such intentional states — an account of intentional realism. And from Alexander's Dictum (and, of course, the consensus in post-Davidson philosophy of action) we can see that establishing an account of intentional realism is going to involve defending the

causal status of intentional properties; in this case, folk psychological states of belief, intention, and the rest.

In short, my account of agency will involve defending the appropriate form of *intentional realism*. As already noted, it will turn out that this account of intentional realism, which is central to understanding agency, is also importantly connected to our capacity for phenomenal consciousness.

Before turning to the main work of establishing these claims, it is worth pausing to compare this view of agency once again with Steward's own view, as considered in the previous chapter. In her view, the concept of agency has at least the following features:

- (i) an agent can move the whole, or at least some parts, of something we are inclined to think of as *its* body;
- (ii) an agent is a centre of some form of subjectivity;
- (iii) an agent is something to which at least some rudimentary types of intentional state (e.g. trying, wanting, perceiving) may be properly attributed;
- (iv) an agent is a settler of matters concerning certain of the movements of its own body [...] i.e. the actions by means of which those movements are effected cannot be

regarded merely as the inevitable consequences of what has gone before.⁵

The project I am concerned with here is establishing the *minimum requirements* for agency: that is, the most basic features of agency, considered within the parameters set by the two limit cases of substance causation and moral agency. In other words, what is needed is a description of the minimum conditions sufficient to distinguish agency from the bare notion of substance causation.

It was one of the faults of the moral agency view, I claimed, that it was too narrowly focused on one area of inquiry (moral responsibility), and ended up simply building the conditions of moral responsibility into the basic concept of agency. Since the notion of moral responsibility is a relatively 'high level' concept, which comes with a lot of theoretical baggage already attached, using this as a touchstone for understanding the basic phenomenon of agency is where the difficulty lies.

In a somewhat similar way, a central feature of Steward's view is *incompatibilism*, itself a relatively high level concept that

⁵ Steward (2012: 71-2). The removed section of the quote is simply a reference to an earlier part of the her book.

comes as part of a whole theoretical package. As such, there is the possibility that the conception of agency thus characterised takes on too many of the theoretical presuppositions of the notion of incompatibilism. Given the more general task being pursued here, of distinguishing agency from non-agency, it might be thought that introducing incompatibilism at this stage makes that view of *agency* weaker, because of the extremely demanding satisfaction conditions for that thesis.

This cannot really be seen as a *criticism* of Steward's view, or of the other views mentioned above, because it does not reflect any particular error or inconsistency in the work. What it does is simply raise the concern that, given this rarefied conception of agency, an opponent might simply argue that, yes, *that* very metaphysically demanding conception of agency may be incompatible with determinism, but it is nonetheless true that *mere agency* is

something that is quite independent of determinism or indeterminism.⁶

In other words, one might object that we can simply conceive of a form of agency that *doesn't* centrally involve metaphysical indeterminism. Precisely because Steward's view has set the benchmark for what counts as agency so high (just like the moral agency view), the response is simply to section off that conception as a 'special case' of the phenomenon in question — or indeed, as a conception of *something else* altogether.

Earlier I rejected the claim on the part of my imagined interlocutor that revising the term 'free will' too far from the folk conception would put me in danger of failing to talk about *free will* at all, thereby invalidating my argument. My response was, more or less, 'so much the worse for the folk conception'. I stand by that claim here, in the case of the agency concept. Rather than *dismiss*

⁶ I think it is clear, of course, that this is exactly Steward's point. Given the way she sets it up, it is not even that there are no deterministic possible worlds in which agency exists — far fewer than that, since many non-deterministic worlds will not contain the right 'kind' of indeterminism to support agency, i.e. it is not the mere falsity of determinism (although that is necessary) that she requires, but a *positive* conception of nondeterministic causation. The bare fact of agency, then, is a very rare phenomenon in logical space, so to speak.

Steward's conception of agency on these grounds, I think that a much better response is this: Steward has articulated *a* concept of agency — or perhaps better, she has articulated the concept of agency*. However, agency* is not in fact the concept that plays a role in our folk psychology in the way I have been considering. Thus, to the objection that we *ought* to accept agency* in this way, in spite of its theoretical commitments, because it is in fact the view that properly characterises our *folk conception* of 'agency', I simply point to the arguments of the previous chapter and deny that it *does* in fact characterise our folk conception.

Of course, one could take exactly the same line with my own view of agency — at least in principle. The difference is that, according to my arguments here, my view of agency does have the support of the 'folk conception'. It is, at least, a *better fit* than the 'agency*' concept for this particular inquiry. The benchmark in each case, as with the discussion of 'free will', is that of determining the *practical* consequences of using one or the other of these concepts.

I claimed earlier that my revision of the term 'free will' makes better sense of various debates in the literature surrounding that topic. That is the practical pay-off for revising in that way. Similarly, the practical pay-off in the case of Vargas's revision of the

conditions for moral responsibility (in favour of compatibilism), is the normative adequacy of the *practises* that follow from accepting that conception of the term.

Setting aside the point about indeterminism, Steward's claim that the agent be properly attributed at least some kind of intentional states, is something that I am clearly in agreement with — as the arguments of this chapter will show. This is perhaps the most obvious similarity between the views, although it is, as I have argued above, a common move in the literature to view the attribution of intentional states as being closely connected to agency. The important question as always is exactly how this is spelled out: the way in which I do so here makes use of a number of features which are not found in Steward's account. With that said, Steward's view appears to share with the present view the thought that it is only when the entity in question is an agent that the attributed intentional states are *genuine*, in a sense in which certain *instrumental* attributions of intentional states are not.

The second requirement found in Steward's account is that the agent be a "centre of some form of subjectivity". As it stands, the notion of 'subjectivity' is too vague to pick out a definite view, and is in any case not an idea that Steward develops in the book

(focusing instead on the arguments for incompatibilism). As such, I take no stand on whether or not the view I develop here is a view that Steward would endorse, or whether it is even similar to a view she might develop: although the notion of ‘subjectivity’ could plausibly be developed in terms of specifically phenomenal consciousness—as I do here—it could also charitably be interpreted in a range of other ways.

Finally, there is a requirement in Steward’s discussion of agency for *self-movement*. I do not include a comparable feature in my discussion of agency. This is because the notion of self-movement can be read in at least two ways. Firstly, it might be nothing more than a quite basic requirement that the agent actually exhibit *observable behaviour*: that there be some behavioural activity which we can interpret as the manifestation of its agency, so to speak. One might contrast this with the case of creatures that were constitutionally incapable of movement, such as Galen Strawson’s imagined example of the Weather Watchers, claiming that such creatures could not be agents because there is nothing that they *do* to identify as agentive activity.⁷

⁷ Strawson (2010: 251).

Of course, one might suggest that the agency of the Weather Watchers is limited to *mental* actions: in the case of mental actions, one could argue that there *is* something that one ‘does’ — cause changes in one’s mental states, and, depending on one’s views, this might (also) be a causing of changes in the body (viz. the brain). The latter appears to be Steward’s own view of the relation between her self-movement requirement and the case of mental actions.⁸ Hence, it is not necessary that there be *observable* movement or change, but only that there be some movement or change in one’s states to identify as agentic activity.

With that said, one could imagine extending Strawson’s example such that the Weather Watchers were not capable of mental “self movement” either. Indeed, Strawson’s own view on mental action suggest that he views the phenomenon as being a lot less common than many philosophers have thought.⁹ In this further imagined case, we might then think that the Weather Watchers were

⁸ Steward (2012: 32-3). She writes: ‘When one actively thinks or undertakes mental operations of any kind, one exercises a power to effect movement and changes in one’s own brain (although one need not be aware that the action has such a description).’

⁹ Strawson (2008: 233).

not agents, because they were not capable of any kind of self movement / change.

On this *weak reading* of the self movement requirement, according to which it is simply the requirement that there be some kind of behaviour that can be identified as the *activity* of the agent, I do not include a comparable feature in my discussion of agency because there is a sense in which it is already guaranteed by the attribution of genuine intentional states such as belief or desire. It is arguably a condition of the genuine attribution of beliefs and desires that there be some range of behaviour that can be interpreted intentionally, even if it is not a *sufficient* condition, as some stronger forms of interpretationism suggest. More generally, it seems almost definitional of ‘agency’ on anyone’s view that there be something that is *done*, whether that is observable behaviour or mental action. If this is all that is required by the self-movement requirement, according to the weak reading, then it seems to add very little to the discussion.

Yet on a *strong reading* of the self movement requirement, the claim is in danger of becoming question begging: since, reading more into this notion of ‘self movement’ leads to a view on which being a ‘self mover’ is tantamount to the claim that one must be an

agent. For example, one might suggest that a vehicle, such as a car, is a 'self mover' because it is capable of moving around 'by itself' (without being pushed or pulled). But no doubt the proponent of the self movement requirement here would reply that this is not a genuine case of *self* movement at all, because the car requires a driver, and this is really just a complex case of efficient causation. But when it comes to spelling out what exactly 'genuine' *self* movement would involve, it is hard to find a condition that is strong enough to say more than the simple requirement that some kind of movement or change occur, but that is not so strong that it amounts to the requirement that there be *agentive activity*; that is, which builds agency directly into a *sui generis* notion of 'self movement'.

I suggest that the weak reading does not need to be included as a stand-alone requirement, because it falls out of the discussion of intentional realism, or it is too general to add anything substantial to the idea of agency. As for the strong reading, I suggest that there is no non-question-begging way of characterising the strong sense of 'self movement' without presupposing agency.

In the remainder of this chapter, I will consider in detail the claim that agency requires intentional realism, say more about what

exactly is involved in such realism, and discuss the connections between that view of intentional states and consciousness. In the following chapter, I will take up in more detail the second set of issues mentioned here, and discuss the relation between perceptual consciousness and actions more generally.

5.3 Challenges to Intentional Realism

I have suggested that our default view of agency commits us to some form of *intentional realism*. But two objections are likely to arise immediately. First, it will be objected that folk psychology is essentially just a primitive ‘proto-theory’, akin to ‘folk physics’, used by people to try and intuitively explain the behaviour of the people around them. By contrast, we now have access to the resources of science, especially cognitive science and neuroscience, to develop a mature theory of human behaviour without relying on the outdated notions of folk psychology. Contrary to some philosophers, it will be argued, there is little prospect of a reduction of folk psychology to some sub-personal level, hence we should abandon use of the former altogether. Call this the *challenge from eliminativism*.¹⁰

¹⁰ The classic sources are Churchland (1981) and Stich (1983).

The second objection can be called the *challenge from instrumentalism*. This view turns on the suggestion that beliefs and desires are ‘useful fictions’ that, although useful enough to count against their wholesale elimination, are not ‘strictly speaking’ real features of the systems to which we apply them. One might argue that, just as we are able to take ‘the intentional stance’ towards a thermostat (for example), this is really no different to what happens when we view other people as ‘believers’ and ‘desirers’. You and I no more ‘have’ beliefs and desires than does the thermostat, it is simply that such locutions add nothing of predictive or explanatory value in the latter case, so there is no advantage to doing so there. But using them in the case of other people adds a lot of *convenience*, and makes explanation a lot simpler.

Both of these objections, if they went through, would undermine the support for my view of agency: they would undermine the case for basing a metaphysical distinction on the putative distinction that is drawn by our folk psychological concepts.

The challenge from eliminativism would directly undermine this distinction, because it denies that there really are any beliefs or desires that could mark any kind of useful distinction. Since the

intentional states themselves do not exist, according to eliminativism, there could of course be no *distinction* between those things that do, and those that do not, have those states (that is, are capable of *being in* those states).¹¹

By contrast, the instrumentalist challenge would undermine the distinction in a less straightforward way. The instrumentalist view makes use of locutions involving ‘beliefs’ and ‘desires’: it will be possible to make predictions about future behaviour, or give explanations for actual behaviour, that invoke these terms. Accordingly, there is a sense in which it will be possible to utilise those terms to mark a distinction between cases. There may be cases in which it is useful, according to some practical purposes, to attribute beliefs and desires to an entity. There will likely be other cases in which there is no advantage to doing so. We might make a distinction on this basis: things in the former class ‘have’ beliefs and desires, and the latter do not.

But the distinction drawn on this basis would not be sufficiently robust. Like any instrumentalism, it depends on the ‘practical purposes’ that set the criteria for usefulness: this line

¹¹ This could of course be put in terms of *properties*, and the possessing or instantiating of such properties.

could be drawn in arbitrarily many places according to the projects of that particular enquirer. Contrary to the instrumentalist view, then, it is necessary to have an objective criterion for the attribution of intentional states in order to support a robust distinction between having or not having genuine intentional states. Hence a defence of intentional realism would constitute a reply to both challenges.

5.3.1 Eliminativism

While it is true that a defence of realism is required, realism about folk psychology is too often associated with what is only one quite specific *form* of intentional realism. On this view, realism about folk psychology entails that it will ultimately be absorbed into a mature science of human behaviour. Fodor, for example, requires that the propositional-attitude states reported by folk psychological explanations turn out to be instantiated by syntactically isomorphic structures in the brain (i.e. the ‘language of thought’). He requires that:

there are mental states whose occurrences and interactions cause behaviour and do so, moreover, in ways that respect (at

least to an approximation) the generalisations of common-sense belief/desire psychology; [and that] these same causally efficacious mental states are semantically evaluable.¹²

Ultimately, Fodor believes that

We have no reason to doubt that it is possible to have a scientific psychology that vindicates commonsense belief/desire explanation.¹³

It is this conception of realism, or something like it, that has been the main target of eliminativist criticism. Note that the challenges to this form of realism have been empirical: the general argument is that a mature science of human behaviour will reveal—and is already revealing—that the empirical commitments of Fodor-style realism are simply false posits. According to Churchland, for example, folk psychology is

¹² Fodor (1994: 5).

¹³ Fodor (1987: 16).

a radically false theory, a theory so fundamentally defective that both the principles and the ontology of that theory will eventually be displaced, rather than smoothly reduced, by completed neuroscience.¹⁴

Hence the eliminativist position shares with this form of realism—what we can call *scientific realism*—a conception of folk psychology as a ‘proto-theory’ of human behaviour, much in the same way as ‘folk physics’ is taken to be a naïve proto-science of physical objects which is straightforwardly replaceable by scientific physics.

The scientific realist about folk psychology often begins with the quite accurate observation that folk psychology tends to get a lot of things right, and serves fairly well as a means of explaining some large portions of human activity. As Fodor puts it, “Commonsense psychology works so well it disappears ...”,

And [it] works not just with people whose psychology you know intimately: your closest friends, say, or the spouse of your bosom. It works with *absolute strangers*; people you wouldn't know if you bumped into them. And it works not

¹⁴ Churchland (1981: 61).

just in laboratory conditions—where you can control the interacting variables—but also, indeed preeminently, in field conditions where all you know about the sources of variance is what commonsense psychology tells you about them.¹⁵

From this, the scientific realist moves to the claim that the justification of folk psychology depends on its being absorbed, in some robust sense, into a mature science of human behaviour.

The eliminativist critics in fact agree with both of these claims: Churchland, for example, agrees firstly that “the average person is able to explain, and even predict, the behaviour of other persons with a facility and success that is remarkable.”¹⁶ Eliminativists also agree with the second claim, that folk psychology is a proto-scientific account of behaviour—a kind of *theory*—and hence that it is in need of absorption into a mature science for its continued justification. Since they believe, however, that this cannot be done, even to a limited extent, they suggest that the theory be entirely displaced, and the posits of that theory

¹⁵ Fodor (1987: 3-4).

¹⁶ Churchland (1981: 62).

(beliefs, desire, and the rest) should be straightforwardly eliminated.

What this discussion of the classic eliminativist debate shows is that the plausibility of eliminativism about folk psychology depends on an unreasonably strong view about the criteria that *realism* about such intentional states would have to satisfy. In the discussion above, what I called ‘scientific realism’ about folk psychology and eliminativism about folk psychology both share the same view about what it would take to vindicate folk psychological explanations of human behaviour.

By contrast, I will argue below for a form of realism about intentional states that is premised on an interventionist account of causation. This sets a quite different benchmark for realism, one that is not open to the same charge of eliminativism: indeed, the difference-making view of causation that is utilised here is the default way of understanding causality in the natural sciences. Far from being empirically implausible, the interventionist method is central to empirical inquiry, especially in the higher-level sciences.

Before turning to develop that view in more detail, I will consider the second of the two challenges noted above: the instrumentalist view. In fact, this is doubly relevant here, because

the instrumentalist view I will consider has been seen as a way of avoiding the eliminativist conclusions noted above, while retaining a ‘kind’ of realism about folk psychology, which is similar to my own approach. Ultimately, however, it does not succeed in this task, and does not secure any kind of realism.

5.3.2 Instrumentalism

Dennett’s ‘intentional systems’ view has been offered as a way to avoid the eliminativism described above, while at the same time resisting the realism that involves accepting something like Fodor’s representational theory of mind. That is, Dennett’s view can be seen as trying to avoid the presumption shared by both the ‘scientific realist’ view, and the corresponding eliminativist view, considered above. This of course is my strategy as well, and so it is worthwhile exploring Dennett’s account to see how far it can take us.

It is a point of contention among many philosophers whether or not Dennett’s view should be considered a form of realism, as Dennett sometimes insists, or simply a straightforward instrumentalism. Dennett continues to maintain that it is a *kind* of realism: a ‘mild realism’, or realism ‘with a grain of salt’.¹⁷

¹⁷ See (1991) for the main defence of realism. For the ‘grain of salt’ see (1988: 73).

Indeed, one must be careful to distinguish instrumentalism from eliminativism about folk psychology: according to some versions of instrumentalism about intentional states, there are no beliefs or desires according to the “ultimate structure of reality”; instead, there is simply the “physical constitution and behaviour of organisms”. What makes this a form of *instrumentalism*, and not eliminativism, is the added proviso that “we are often well advised to tolerate the idioms of propositional attitude” for practical purposes in certain contexts.¹⁸

The difficulty is that there is a tension between denying that there “really are” such things as beliefs and desires, on the one hand, and maintaining that the framework of folk psychology nonetheless has a legitimate practical use, on the other. For if there really are no such things, then it is hard to see how making use of them would confer any *advantage*, in practical contexts or anywhere else. The more that one emphasises the practical value of folk psychology, and the extent to which it works as a predictive tool, for instance, the more that the claim about the non-existence of such things as beliefs and desires is problematised. In fact, Fodor makes a similar point:

¹⁸ These three quotations are from Quine (2013: 202).

After all, we do use commonsense psychological generalizations to predict one another's behaviour; and the predictions do—very often—come out true. But how could that be so if the generalizations that we base the predictions on are *empty*.¹⁹

Dennett's claims to realism notwithstanding, his view can be seen as an attempt to accommodate something like this basic insight, i.e. that the practical adequacy of folk psychology counts against its outright elimination.

Dennett compares the question whether we ought to be realists about the beliefs and desires posited by folk psychology to the question whether we should 'be realists about' the notion of a *centre of gravity*. His answer is that we should be realists about belief in the same way that we are realists about centres of gravity. Beliefs and desires, he claims, are like *that* — a kind of abstract entity. But what kind of realism is this? Dennett notes that there are two ways of approaching that question.

¹⁹ Fodor (1987: 4).

First, the ‘metaphysical path’, according to which we are concerned with “the reality or existence of abstract objects generally”.²⁰ In that sense, beliefs and desires, like centres of gravity (and any number of arbitrarily defined abstract objects, such as the ‘centre of population of the United States’), have the same metaphysical status as all abstract objects, such as numbers, the empty set, and the novels of Jane Austin.

That is as much as Dennett says about the ‘metaphysical status’ of beliefs and desires in this context, presumably taking them to be no more or less problematic, metaphysically speaking, than the existence of abstract objects in general. By contrast, he is much more interested in the “scientific path” to answering the question of realism. On this view, centres of gravity are “good” abstract objects,

They deserve to be taken seriously, learned about, used. If we go so far as to distinguish them as *real* (contrasting them, perhaps, with those abstract objects which are *bogus*), that is

²⁰ Dennett (1991: 28).

because we think they serve in perspicuous representations of real forces, "natural" properties, and the like.²¹

This way of characterising the 'reality' of abstract objects, he claims, gets us closer to what he would like to say about the reality of beliefs and desires. What is 'good' about some abstract objects is their *usefulness* in certain applications — once again, a broadly instrumentalist strategy. The notion of a 'good' abstract object is a basically normative one fixed by the practical requirements and projects of a particular enquirer: in this case, one engaged in the project of certain empirical calculations that can be facilitated by utilising the notion of a *centre of gravity*.

But Dennett insists that they are useful *because* they (can be *used* to) track certain real features of the world: real forces, as it may be. This second point is meant to establish the 'kind of' realism that Dennett insists is delivered by his account, and which makes it more than a mere instrumentalism. How does this analogy with centres of gravity apply to folk psychology, that is, to the reality of beliefs and desires? Dennett elaborates on the above claims in his discussion of *real patterns*. In this case, too, the abstracta of beliefs

²¹ Dennett (1991: 29).

and desires track some real feature of reality, even though such 'entities' themselves are, strictly speaking, no more to be found in concrete reality than are centres of gravity.

The features of the world that they track are intentional 'patterns'. Dennett suggests that the patterns which are tracked by the abstracta of 'belief' and 'desire' are "discernible in agents' (observable) behavior when we subject it to 'radical interpretation' [...] 'from the intentional stance'".²² Dennett's view ties the reality of beliefs and desires to the existence of an *interpreter*; that is, to the existence, at least in principle, of an observer who is able to interpret the outward behaviour of the individual according to 'the intentional strategy'.

Hence, Dennett is committed to at least some form of interpretationism. Indeed, he defines the notion of a pattern itself in terms of the possibility of pattern *recognition*.²³ Thus the ambiguity in Dennett between realism and instrumentalism is recapitulated once again in the notion of a real pattern: the pattern is *there* to be

²² Dennett (1991: 30). The reference to 'radical interpretation' is a nod to Davidson, who shares a somewhat similar view to that of Dennett on the question of folk psychology.

²³ He writes that "in the root case a pattern is 'by definition' a candidate for pattern *recognition*" (1991: 32).

seen; but it is only relevant, and only has a function, when it is *seen*.

Dennett writes:

These patterns are objective—they are *there* to be detected—but from our point of view they are not *out there* independent of us, since they are patterns composed partly of our ‘subjective’ reactions to what is out there, they are the patterns made to order for our narcissistic concerns.²⁴

He compares this description to a comment by Anscombe, who writes of “an order which is there whenever actions are done with intentions”,²⁵ suggesting that she may have been making a similar point. He elaborates:

If you “look at” the world in the right way, the patterns are obvious. If you look at (or describe) the world in any other way, they are, in general, invisible.²⁶

²⁴ Dennett (1987: 39).

²⁵ Anscombe (1963: 80).

²⁶ Dennett (1987: 39, n. 1).

The essential connection to the perspective of an (in principle) observer is something that is familiar from interpretationist accounts of the mind. However, Dennett maintains that the patterns which are picked out by observers occupying the intentional stance are *abstractions* from the activity of many lower-level ‘minutiae’ which produce the pattern visible in observable behaviour. What the intentional stance provides is “computational leverage” over the lower-level (physical) details, by providing predictions at a “scale of compression” that is vastly greater than the raw “bitmap” — i.e. the physical-level description of those same events.²⁷

This is not to suggest that Dennett believes, after all, that the patterns picked out from the intentional stance are to be found, like Fodor suggests, in the workings of the brain. But the emphasis on ‘abstraction’ and ‘compression’ from the ‘bitmap’, and similar comments, suggests that Dennett has a view of the relation between the level of folk psychology and the level of sub-personal mechanism (for example, the workings of the brain) according to which facts at the former level can be—in principle—explained again at the lower level.

²⁷ Dennett (1991: 42-3).

Or more precisely, that they can be *explained already* at the level of the most basic physical constituents—in principle, of course—without ever mentioning anything intentional. Such a project would no doubt be impossible for finite creatures like us, not least because of whatever constraints there may be on the greatest physically-possible amount of computational power available in the universe: we get by with the *approximations* of folk psychology, which compress the enormity of the physical data into useful abstractions.

This characterisation of the relation between the level of folk psychology (the “personal level”) and the “sub-personal” level is such that it puts into question the kind of intentional realism that can be extracted from Dennett’s view. It is true that the existence of intentional patterns is not *merely* observer-relative, because those patterns are objectively available for any possible observer to detect. Yet those patterns which are detected by observers via the intentional stance are, in the end, simply *approximations* or, as Dennett puts it, computational shortcuts, to the sub-personal (ultimately physical-level) truth of the matter.

So far I have suggested that Dennett’s view is best seen as a novel form of instrumentalism, which is targeted at resisting the

slide into eliminativism that I cautioned against earlier. Instrumentalism, as I have already suggested, is insufficient for present purposes: what is needed is a stronger thesis, one that is *straightforwardly* realist about intentional states. Yet, Dennett's view can help with this project: it contains the resources for constructing a more robust view of the mind than Dennett's own view provides.

Whether or not it is true that the view I am proposing is Dennett's own view, there certainly is *a* view which can be developed along these lines. There are points at which Dennett makes suggestions that are indicative of the kind of difference-making approach that I am arguing we need in order to support intentional realism, and parts of his view can certainly be read profitably along those lines. Maybe, in the end, Dennett *is* a kind of intentional realist—the kind that I am about to describe—and it is merely his own reluctance to be pigeon-holed by the term 'realist' that prevents him from stating things in this way.

5.4. Interventionism and Mental Causation

Based on the discussion above, one might object that Dennett's view of the relation between the personal and the sub-personal levels undermines the explanations of behaviour which are given

from the intentional stance (the personal level). In fact, there are points at which Dennett appears to address this challenge directly. In a footnote to his paper on real patterns, he considers something like this objection:

Several interpreters of a draft of this article have supposed that the conclusion I am urging here is that beliefs (or their contents) are *epiphenomena* having no causal powers

Interestingly, Dennett's response to this objection is to note that

If one finds a predictive pattern of the sort just described one has *ipso facto* discovered a causal power — a difference in the world that makes a subsequent difference testable by standard empirical methods of variable manipulation. Consider the crowd-drawing power of a sign reading "Free Lunch" placed in the window of a restaurant, and compare its power in a restaurant in New York to its power in a restaurant in Tokyo.²⁸

²⁸ Dennett (1991: 43, n. 22).

Dennett's point appears to be that the explanation of the sign's power to draw a crowd in New York is no less a causal explanation simply because it is stated in intentional terms. He argues that it would only be because a philosopher was relying on a "pinched [notion] of causality derived from exclusive attention to a few examples drawn from physics and chemistry"²⁹ that she might think the above explanation could not be a causal explanation.

Now it is quite true that the intentional explanation Dennett gives here ought to be considered a causal explanation: indeed, it is a central part of realism about the mind that such intentional explanations be considered causal explanations, as the discussion of Alexander's Dictum above has made clear.

However, the present objection was that Dennett's view cannot support realism about the intentional explanations that he claims are visible 'from the intentional stance', and this objection really has two parts. To the objection that intentional stance explanations are not *causal* explanations, Dennett's response is surely right: there is no reason to think that such explanations cannot be causal, simply because they are stated in intentional terms.

²⁹ Dennett (1991: 42, n. 22).

The second aspect to the objection, however, can be pressed further: given Dennett's views on the relation between the intentional level and the lower levels (or stances one might take), why is the causal explanation discerned at the intentional level not *undermined* by explanations at lower levels. What gives priority to explanations framed in terms of intentional properties, and not, for example, in terms of neuroscience? Or in terms of whatever basic physical processes can be invoked to describe the event?

The remarks in the quotation cited above suggest a promising way of answering these questions: Dennett notes that if one finds a 'predictive pattern' of a certain kind, then one has ipso facto found a causal power: that is, "a difference in the world that makes a subsequent difference testable by standard empirical methods of variable manipulation". The best way to understand these remarks is according to a difference-making or manipulationist view of causation — specifically, I suggest, an *interventionist* view of causation.

The question I raised above was whether there is a way, on Dennett's view, to explain why we should consider the properties on view from the intentional stance to be causes in their own right. For Dennett, those properties are defined by at least partly

pragmatic concerns, as a means to dealing with physical systems that are simply too complex to interact with at the physical or biological level.

Notwithstanding Dennett's suggestive comments about manipulationism, the difficulty is that it seems that on Dennett's view we merely have to 'get by' with the computational shortcuts afforded by abstractions at the intentional level, but that there remains a sense in which we should like to 'really' explain things at the biological, or even physical, level — were that to become possible for us. Hence, the intentional explanations are not sufficiently 'robust', in a sense I will make clear in this section.

At this point we can pick up on the nod towards manipulationist views of causation, and apply them to our present concerns. John Campbell takes a broadly manipulationist approach to the problem of mental causation: in particular, he makes use of a theoretical device he calls a 'control variable' in his discussion the issue. The notion of a 'control variable' will provide a *principled* way of establishing causal relations that involve specifically intentional properties in their own right, rather than as a merely pragmatic means of bypassing the unwieldy physical calculations that would otherwise be required.

The view that Campbell is working with is a form of manipulationism, but he is working within a specific form of that general view which is called *interventionism*. Interventionist views are a special case of the general manipulationist approach to causation. According to interventionist views, for X to be a cause of Y is for intervening on X to be a way of intervening on Y.³⁰ The basic intuition behind this approach to causation is that causes are potential ways of manipulating their effects.³¹ What specifically *causal* information adds to information about correlation is therefore information about *manipulability*.³² Correlations do not have this feature: if X and Y are merely correlated, then it is not the case that manipulating X is a way of affecting Y.

In contrast to this way of formulating the ‘difference making’ part of the view, earlier examples of manipulationism leaned much more heavily on the notion of a ‘free action’ in formulating their account: these are often called ‘agency theories’ of causation for this reason.³³ The basic idea there was to give a non-circular, reductive account of causation. The notion of a ‘free action’ used in

³⁰ Campbell (2007: 59).

³¹ Woodward (2007: 20).

³² Strevens (2007: 233-4).

³³ See for example von Wright (1971) and more recently Menzies and Price (1993).

characterising the manipulation, although it appears to presuppose the idea of causation, was supposed to depend on our directly experiencing our own activity as agents: it was claimed that we can grasp a kind of ostensive definition of a basic notion of “bringing about”, which can then provide the ground for a reductive analysis of causation.

The purpose of the present inquiry is not to develop an analysis of causation, reductive or otherwise, but I do utilise a manipulability view of causation in an important way, so it is worth pointing out some of the problems with the agency view. Furthermore, I am concerned with the concept of *agency*, and not directly with the concept of *causation*: this is another reason why the aspirations of the ‘agency’ type of manipulationism do not fit well with my own purposes here, since they *start* with the notion of a ‘free action’ already taken for granted.

Much of the criticism of ‘agency manipulationism’ has been focused on (i) the notion of a ‘free action’ and its role in providing a reductive analysis of causation, and (ii) the threat of circularity in such manipulability accounts.

The interventionist view can be seen as an improvement over the agency theories with respect to these difficulties, and for that

reason I choose to work within the general interventionist approach in what follows. The basic idea that the agency theory tries to capture with the notion of ‘bringing about *X* by a free act’, when it is properly refined and clarified, is more or less what the notion of an intervention is designed to do. Indeed, the notion of a ‘surgical’ intervention is precisely a way of making sufficiently clear what kind of manipulation is required to reliably discover a causal relationship between two variables, something that was arguably under developed on the agency views.

Secondly, the interventionist view dispenses with the reductive aspirations of the agency theory of causation. The worry there was that the notion of a ‘free act’ directly involves the idea of causally affecting something, and, notwithstanding the comments about a direct experience of the basic phenomenon of ‘bringing about’, this renders the account circular. Since the agency view claimed to offer a reduction of causal talk to non-causal talk, presupposing causation in this way is viciously circular. By contrast, the interventionist view does not claim to be reductive, but arguably it *is* non-circular and non-trivial in an important way.

It is non-circular because, as interventionists such as Woodward have argued, in unpacking that claim that *X* is a cause

of Y , we might appeal to the notion of some process I as being an intervention on X . But for the purposes of characterising the intervention I , we would not need to make use of any information about any causal relationship that may or may not obtain *between* X and Y . As Woodward puts it:

we may use one set of claims about causal relationships (e.g., that X has been changed by a process that meets the conditions for an intervention) together with correlational information (that X and Y remain correlated under this change) to characterize what it is for a different relationship (the relationship between X and Y) to be causal.³⁴

Hence, it is not a reductive analysis of causation, but neither is it viciously circular in the sense that it presupposes information about the very causal relation that one is trying to characterise.

Furthermore, it is non-trivial because the interventionist view could conflict with other views about causation in its verdicts about particular cases. For example, the interventionist approach

³⁴ Woodward (2013: §7). See also Woodward (2003) for the full defence of these claims.

returns a positive verdict on the possibility of causation by omission (failing to water the plant caused it to die) — in direct contrast with (e.g.) causal processes views of causation, such as that given by Salmon or Dowe.³⁵ The possibility of inconsistent verdicts on these cases by interventionist views and non-interventionist views means that the interventionist view is not vacuous or trivial.

There have been several attempts to apply the general interventionist (or manipulationist) approach to the problem of mental causation. I said above that one interesting development due to Campbell is the notion of a control variable: I will now consider Campbell's view in some detail, before briefly comparing his conclusions with some similar views in the literature.

5.4.1 Causation and Control Variables

To understand how the idea of a control variable can help with the problem being considered in this section, consider one of Campbell's own examples. Imagine that you have an ordinary radio and you want to know what causes the volume of the sound coming from the radio.³⁶ Also imagine that you have access to all

³⁵ Woodward (2013: §7). The process views are Salmon (1984) and Dowe (2000).

³⁶ Campbell (2010: 22-3).

possible information about the radio, including all the physical level details. Since you in fact have all the information about the radio available to you, why *not* seek the explanation at the physical level? In the discussion of Dennett's view above, it seemed that the only reason not to 'descend' to lower, non-intentional levels of explanation was computational intractability. That, we are now imagining, is not the case with the radio.

Campbell imagines that we might specify a physical level variable that indexes the state of each physical particle at any one time, in order to represent the total state of the radio, at a given time, at the physical level.³⁷ This is to be the putative 'cause' variable. Now, it will be true that (some) interventions on this variable make a difference to the volume output of the radio. Does this mean that we have found the cause of the volume? Not so, according to Campbell, because the supposed cause variable here is not sufficiently 'good' or 'systematic': for a particular variable to be good or systematic in this way it must meet several requirements, and the enormously complex physical variable just proposed fails

³⁷ Assuming for the example a simplified view of physics. Assuming real-world physics would only support Campbell's argument, since it would render even more implausible the suggestion that such a massive physical variable could be in some sense the 'cause' of the outcome in which we were interested.

these criteria, therefore failing as a putative cause of the volume output.

Firstly, it should be that the function from cause variable (X) to effect variable (Y) is *total*, in that every value of X is mapped to some variable of Y .³⁸ This is because while some interventions on the microphysical state of the radio will affect the volume, some will make no difference at all — and some will result in the destruction of the entire radio. Indeed, it is plausible that *many* of the individual interventions that are possible at the microphysical level will have no discernible effect on the output volume, or indeed on any ‘macro’ level variable.

Secondly, Campbell proposes that the relation between these variables should not only be total, but that it should also exhibit a *dose-response relationship*.³⁹ He notes that one of the criteria for causation famously outlined by the epidemiologist Austin Bradford Hill was that a dose-response relationship of this kind strengthens the case for establishing a causal relation between the two variables: if you smoke more cigarettes, for example, you are more likely to get cancer.

³⁸ Campbell (2010: 23-4).

³⁹ Campbell (2010: 25).

While the existence of a dose-response relationship of this kind is commonly recognised as being important in the understanding of causation, it has not always been clear exactly why it is so important. Campbell's own suggestion is that the dose-response relationship is important because it establishes that the variable in question is at the 'right level' to be a cause of the outcome that we are interested in. And since there would not appear to be any kind of dose-response relationship between our gerrymandered physical variable and the output volume of the radio, this suggests that what is wrong with the putative physical cause is that it is at the wrong level to be a cause of the outcome in which we are interested: the volume of the radio.

How does this example carry over to the case of intentional states and behaviour? The intended point of Campbell's analogy with the radio is that in many cases the proper 'control panel' for manipulating the outcome that we are interested in (actions) is going to be found at the psychological or intentional level of agents' mental states. He writes:

The point is almost too obvious to spell out, but people's behaviours vary systematically with their psychological states.

The greater my degree of concern about my missing dog, the more assiduously I will look for her. The more focused I am on a mathematical problem, the harder I will be to distract.⁴⁰

Campbell's point is that there are going to be many cases in which we find intentional states that meet the requirements noted above, and hence should be seen as plausible candidates for the causes of whatever behaviour we are interested in. It is not simply that some changes in an agent's mental states are correlated under interventions on those states with some change in behaviour: in many cases, there will be a *systematic* relationship between these variables, involving *total mapping* of cause variables (changes in mental states) with the effect variables (changes in behaviour); and a *dose-response* relationship, as Campbell highlights in the extract above.

This much serves to establish the plausibility of an interventionist treatment of mental causation, but such a view is likely to encounter several common objections: (1) it might be thought that there is a worry about 'competition' with non-mental causes of the same behaviour, such that the physical cause

⁴⁰ Campbell (2010: 26).

‘excludes’ the putative mental cause; or (2) one might worry that mental states are simply at ‘the wrong level’ to be causes of behaviour, and that there *must* be some physical cause to be found at a lower level; finally (3), it might be objected that accepting mental states as causes in this way would be at odds with various naturalisation projects, and a general scientifically-informed worldview.

In response to (1), one can note that the interventionist view of causation is well placed to handle such ‘exclusion’ arguments against mental causation. In particular, work by Peter Menzies and Christian List directly addresses such ‘exclusion arguments’, and draws the interesting conclusion that, while it is true that there are times when mental states may be ‘excluded’ by lower-level physical states—as in the standard exclusion argument—there are also many cases in which *physical* states are excluded as causes by higher-level mental states.⁴¹

This turns out to be a basically contingent matter, depending on the details of the particular case. So, while they repudiate the

⁴¹ List and Menzies (2009). See also Menzies (2013), in which Menzies also summarises work from List and Menzies (2009). In the text, I use the name “Menzies and List” when referring to either piece of work.

claim that there cannot be any mental causes of behaviour, neither do they suggest the opposite general claim and suggest that there is always going to be a mental cause of some behaviour.

Central to their argument that mental states are not always excluded by lower level physical states is a difference-making account of causation. The basic claim is that exclusion arguments against mental causation mistake causal sufficiency for *causation*. In general, such arguments claim that where a mental state *M* is a cause of behaviour *B*, and where *M* is realised by physical state *P*, the mental state is excluded by *P*, because *P* is *causally sufficient* for *B*. But Menzies and List argue that causal sufficiency is not the same as causation: they cite Wesley Salmon's observation that, while a man's taking a contraceptive pill is causally sufficient for his not getting pregnant, this is not causation because the man's taking the pill *makes no difference* to his not getting pregnant.⁴²

The main reason for this is that citing the man's taking the contraceptive pill is *overly specific* as a putative cause of his not getting pregnant. On this point, we can cite Stephen Yablo's familiar example of the pigeon that is trained to peck at all and only

⁴² Salmon (1971: 34). Cited at Menzies (2013: 72).

red objects.⁴³ In this example, the pigeon is shown a red object that it proceeds to peck at. In this particular case, the red object happened to be *scarlet*. What is the cause of the pecking? Is it the fact that the object was red, or the fact that it was scarlet?

According to the kind of ‘exclusion argument’ reasoning that we are considering here, the object’s being *red* in this case is realised by its being *scarlet*, and for that reason being scarlet is causally sufficient for the pecking — and so *is the cause* of the pecking. But, as Yablo points out, citing the fact that the object is scarlet as the putative cause is overly specific: the pigeon would have pecked if it was any shade of red, but if it had been anything other than *red* the pigeon would not have pecked. What makes the difference to the effect (pecking), is whether the object is red or not red, and hence that is the right level at which to cite the cause of the pecking.

Applying these considerations to mental causation, Menzies and List report that there are many cases in which citing a mental state as the cause will be *proportional* to the effect in question (some behaviour), while the putative physical state is not, and that the mental state therefore *excludes* the physical state as the proper cause of the effect.

⁴³ Yablo (1992). Cited at Menzies (2013: 72-3).

For example, consider a straightforward case in which a mental state M is realised by a physical state N , but where M is *realisation-insensitive*. That is, according to the familiar multiple realisability of mental states, M could in fact be realised by a range of similar physical states $\{N_{i1}, N_{i2} \dots N_{in}\}$. On this occasion, suppose M is realised by N_{i1} . The question, as above, is whether we should say that M is the cause of some behaviour B , or whether N_{i1} is the cause of B . And just like the example given by Yablo, we will find that M is proportional to the effect in question: whether or not the subject is in mental state M is what makes the difference to whether or not B occurs. If the subject was not in physical state N_{i1} , the closest possible world in which that is true is a world in which the subject is in physical state N_{i2} , meaning that B would still occur. Yet the closest world in which M does not occur is one in which some *other* mental state occurs, and hence makes a difference to the behaviour. Hence M is the cause of the behaviour B , and N_{i1} is not.

In response to the second objection (2), Campbell considers the suggestion that psychological variables cannot properly be causes because they are simply at ‘the wrong level’. One common way of developing this objection is to suggest that only variables that figure in the characterisation of *mechanisms* are causally

significant, and there are no definitively psychological mechanisms — in contrast to, for example, physical or biological mechanisms. In reply, Campbell points out that causation is not defined in terms of mechanisms, but in terms of what happens under interventions: in short, causation is one thing, and the mechanism by which it happens is another.⁴⁴

Indeed, if this ‘wrong level’ argument were successful, the claim that ‘smoking causes cancer’ would face the same objection: there is no doubt some biological mechanism by which smoking causes cancer, but the term ‘smoking’ is not a biological notion. It is not “a term defined at the level of mechanism”.⁴⁵ It would be implausible to suggest, on this basis, that smoking does not cause cancer. So too for psychological variables:

there may be a brain mechanism by which grief causes anger; ‘grief’ and ‘anger’ are not themselves terms defined at the level of the mechanism, but that does not show that they are not causally significant.⁴⁶

⁴⁴ Campbell (2010: 18-19).

⁴⁵ Campbell (2010: 20).

⁴⁶ Campbell (2010: 20).

Another way of developing the objection that psychological variables are ‘at the wrong level’ to be causes is to suggest that only properties that are governed by *exceptionless general laws*. There are two responses to this.

Firstly, as Campbell points out in his discussion, the whole methodology of the randomised controlled trial is premised on the idea that there are not going to be exceptionless general laws at the level of the variables that are being investigated, in medicine, and in all of the non-fundamental sciences. As he puts it

Anyone who takes a medicine with the idea that it is going to do something to them is abandoning the notion that exceptionless general laws are required for causal connections.⁴⁷

But there is a deeper point to make here. An unspoken assumption in these objections is that lower level physical causes are somehow more ‘real’, or that it is at the physical level where the real causal action happens, and higher level causes—if there are such things at all—are somehow parasitic on that. Yet it is by now a familiar point

⁴⁷ Campbell (2010: 20).

that the science which studies the *most* basic, or fundamental, physical level actually seems to have little use for the ordinary notion of causation.

Indeed, the interventionist view of causation straightforwardly accommodates this point: the whole methodology of interventionism is such that we can identify causation only when we have delineated a well defined subsystem, and are able to identify, in relation to that subsystem, a set of ‘exogenous’ variables in terms of which we can define an intervention on that system. Fundamental physical theories, for example, are such that they apply to the whole universe — but this feature ensures that they are *not* in fact construed causally, on the interventionist analysis.

Campbell points out that ‘the objective of finding a comprehensive scheme of variables that characterizes everything that happens is similarly antithetical to the idea of causal connection.’⁴⁸ Far from *requiring* the existence of universal exceptionless laws, then, the task of discovering causal connections is a quite different project. Pearl sums up this thought:

⁴⁸ Campbell (2010: 21).

If you wish to include the whole universe in the model, causality disappears because interventions disappear – the manipulator and manipulated lose their distinction.⁴⁹

Finally, objection (3), concerning the more general worry about a naturalistic vision of the world. Does accepting psychological states as causes stand in conflict with this worldview? The interventionist view of causation is primarily an exercise in methodology, not metaphysics: simply put, it tells you how to find out whether, and which, phenomena are causally connected. Woodward is clear that

one of the attractions of the manipulationist account is precisely its unmetaphysical character — rather than thinking of causal relationships as involving mysterious other worldly entities [...] I urged instead that we think of them simply as relationships that are exploitable for purposes of manipulation and control. [...] For those who care about metaphysics, this

⁴⁹ Pearl (2000: 350).

sort of view might be supplemented by any one of a number of different stories about metaphysical foundations.⁵⁰

Any number of ‘metaphysical foundations’ are compatible with the interventionist (manipulationist) conclusions about the causal reality of mental states. The minimum that is required to satisfy a general naturalistic view of the world is that such mental states supervene on the physical, which is perfectly compatible with Woodward’s comments on the metaphysical commitments of interventionism, and with the conclusions of the interventionist approach to causation.

Consider briefly a thought experiment proposed by William Seager. Interestingly, Seager uses this thought experiment as part of a discussion of Dennett’s views on the naturalisation of the mind. Seager asks us to imagine programming a computer simulation of some simple physical phenomenon, such as a pendulum swinging on the surface of the moon.⁵¹ The simulation only represents this limited region of space and time (assume that there are boundary

⁵⁰ Woodward (2008: 194) — which is a response to Strevens (2007), cited above.

⁵¹ Seager (2010). The following description of the thought experiment is drawn from (116-7).

conditions to represent external influence). Importantly, the programmer is not permitted to utilise any ‘gross parameters’ such as the *mass* or the *length* of the pendulum, for example, but must program the simulation entirely in terms of the most basic physics. Thus the ‘programming language’ for this simulation is one that only makes use of the terms of the most fundamental physics.

Seager now asks us to imagine a more complex case: a simulation of a child’s birthday party. Again, the programmer is only allowed to utilise the ‘programming language’ of physics. The point of the thought experiment is this: if you believe that what this simulation will show will ‘agree with reality’ then you believe in what Seager calls the Scientific Picture of the World. In other words, you believe that everything supervenes on the physical.

I take it that this thought experiment allows an intuitive grasp of what supervenience on the physical amounts to. Dennett no doubt believes in this view of the world, at least in some sense. And Campbell too would surely agree to this rather general claim. The point here is that everything might supervene on the physical level, in the way that Seager’s thought experiment brings out, but in the end this says little about the *causal* structure of the world,

leaving it open for the interventionist to draw the conclusions noted above.

Far from being in tension with a naturalistic view—or ‘Scientific Picture’—of the world, the interventionist’s conclusion regarding the causal reality of intentional states actually supplements that general metaphysical picture of the world, filling in (causal) details that are simply left open by the claim of physical supervenience.

5.5 Conclusion

The commonsense view of agency is committed to a realism about intentional states. The two principal challenges to intentional realism *itself* come from eliminativism and instrumentalism, and both of them turn on the notion of causation. Specifically, the problem is to establish how such intentional states or properties are causal integrated into the physical world (viz. behaviour).

The interventionist view of causation proved to be the most plausible way of responding to these challenges, while maintaining a genuinely realist view of the intentional causation of behaviour. However, as we shall see in the following chapter, the defence of intentional realism is not yet complete: a central part of this account

requires not just the involvement of mental states in the way so far discussed, but the contribution of *conscious* mental states. Specifically, it will require the input of phenomenal consciousness.

6

Consciousness in Action

6.1. Introduction

Here is the state of the argument so far. Chapter 2 set out the general argument for revising our approach to the free will problem, and suggested instead the adoption of two more specific philosophical projects. One of those projects is taken up as the central part of this thesis: an account of agency that is independently plausible, and that has been doing unnoticed work in the background in the literature on 'the free will problem'. Chapter 3 then set out the motivation for that agency project, and showed how the notion of 'agential control'—which is really a constitutive part of agency itself—has in fact been obscured by the dialectic of the 'free will debate'.

Chapter 4 took up the challenge of getting a grip on the concept of agency itself, given the impressive range of uses to which that term has come to be put. I suggested that Helen

Steward's account of agency has several features which make it a plausible starting point, and then took up an important question raised by Steward's project in that work: what can we learn about agency from investigating our own psychology, in particular, by looking at the 'folk psychological' concepts we apply to certain features of our experience?

Building on that starting point, I proposed a nonreductive account of agency that in fact differs in substantial ways from Steward's own view. Two of the central claims of my view are (i) the commitment to intentional realism, and (ii) the involvement of perceptual consciousness in action control. Importantly, I claimed that these are not arbitrarily connected features, but that they come together as a 'package', and understanding this account of agency requires getting clear on the interconnections between these phenomena. Indeed, a central part of the argument is that defending (ii) in the way that I do here is actually a necessary part of (i). So these are not separate tasks, but simply different stages in the same argument.

The previous chapter took on intentional realism. The conclusion there was that reference to intentional properties plays a central role in action explanation, and that we are specifically

committed to the *causal* reality of such phenomena. I picked up on certain suggestive remarks in Dennett's account, and developed them along the lines of an interventionist account of causation: I claimed that this is the best way to defend a realist view that falls between the extremes of instrumentalism about intentional properties, on the one hand, and the 'industrial strength' realism of Fodor and others, on the other hand. The interventionist's view of causation is, contrary to certain arguments, sufficient to establish the causal reality of mental states. But the conclusion of that chapter established only that this account of intentional realism is a *plausible* one, and that it makes sense of the requirements for agency that I set out in that chapter. The further task is to establish that it is *true* in our own case.

Hence we arrive at the present chapter: since one of the main claims of the view elaborated in the previous chapter is that realism about intentional states requires the possibility of an *intervention* on those states, the argument of this chapter is that consciousness can—and actually does—function as an intervention on an agent's mental states. In particular, we should find that the intentional level is the appropriate (proportional) level at which to locate the 'control panel' for the outcomes we are interested in — intentional actions.

Note that this defence of intentional realism requires that consciousness functions as an intervention on an agent's mental states, *and* that the content of conscious perception is utilised in the control of intentional action. Establishing a robust form of intentional realism, in the way that the previous chapter suggested was necessary, depends on both of these tasks. Recall the discussion of Alexander's Dictum: it was a criterion of the genuine reality of intentional states that they have causal effects — 'to be real is to have causal powers'.

Showing merely that an agent's intentional states can be affected by conscious perception is insufficient: it needs to also be shown that such states *make a difference to behaviour*, otherwise there is an important sense in which such intentional states are epiphenomenal, and this counts against the kind of realism that was proposed in the previous chapter. Hence it is a main concern of this chapter to defend the present view against the charge of epiphenomenalism, in order to complete the argument for intentional realism that was initiated in Chapter 4.

6.2 Intervening on Beliefs

Perhaps the most straightforward sense in which consciousness acts as an intervention on an agent's mental states is in the case of belief — specifically, beliefs about the agent's immediate environment. Now the purpose of an intervention, according to this view of causation, is to isolate a variable of interest and manipulate that variable in a systematic way in order to study the effects of that intervention on a certain phenomenon of interest (the effect variable). In practice, this is usually characterised by an experimenter's intervention on a variable in the context of a controlled experiment, as in a randomised controlled trial.

Put simply, the experimenter takes control of the variable X , and 'tweaks' it in some way (causes a change in the value of the variable), and looks to see what happens to the effect variable Y — the important feature here is that the intervention be 'surgical'. What this means is that the intervention on X should be such that all other causal influences on X are suspended; and that, importantly, any change in Y that occurs as a result of the

intervention on X should come about *only* via the causal connection (if there is a causal connection) between X and Y.¹

This way of characterising an ‘intervention’ is very much tied to the notion of an *experimenter*, carrying out interventions on a range of phenomena in the context of a scientific experiment. But it might be thought that this characterisation fails to apply to the way that we form beliefs: there is no experimenter manipulating the features of our experience, or if there is, then this cannot inform us about the normal processes of belief-formation.

However, the interventionist view can happily accommodate the notion of an intervention in which no such ‘experimenters’ are involved: there just has to be the right kind of natural process, with the right kind of causal history.² In science, this is sometimes called a ‘natural experiment’: as when, for example, it would be unethical or impractical to carry out a controlled trial to study the effects of certain phenomena. A natural experiment of this kind requires that there be some people who are subject to some outside influence (the ‘intervention’), and that it is possible to compare them to a second population who are as similar as possible, but for the fact that the

¹ Woodward (2003: 94, 97). See also the ‘arrow breaking’ conception of interventions, found in Pearl (2000).

² Woodward (2003: 94).

second population is not subject to the ‘intervention’. A common example is that of twin studies, where children who are genetically identical are brought up separately in different conditions, or are subject to differing external influences — allowing the study of the causal significance of environmental risk factors, while controlling for the influence of genetic factors.³

The ‘natural experiment’ view of intervention offers a more promising way of thinking about the role of perceptual consciousness in belief formation.⁴ In the most basic case, conscious perception of some feature of your environment is an intervention on your beliefs about the world: specifically, of your beliefs about that portion of the world that you are perceiving.

This can happen passively, but there are many cases in which the subject consciously *attends* to some feature of the world for the purposes of ‘setting up’ such a natural experiment. For example, suppose that you want to know the colour of a particular object, so you fix your attention on the object in question, and attend in

³ Campbell mentions this example at (2010: 18).

⁴ Although both characterisations are just that—characterisations—and are not *separate* forms of interventionism. They are simply different ways of ‘picturing’ the account of causal inference: by thinking of a randomised controlled trial, or a natural experiment.

particular to that dimension of the object (its colour properties). You therefore act to make it the case that the world itself acts as a *control variable* for your beliefs about the colour of the object.⁵

Consider an example originally presented by Austin.⁶ Suppose you have heard that there is a pig roaming these woods, although you think it is fairly unlikely. You then come across one in a clearing, and look at it in good light; you walk around it, prod it with a stick, sniff it a few times. The pig in this situation is having a causal impact on your beliefs. Importantly, it is not simply that the pig *influences* your beliefs, but that it has a *decisive impact* on your beliefs, viz. about whether or not there is a pig roaming these woods.

⁵ Campbell (2010: 28-9).

⁶ Austin (1962: 114-15). Campbell develops this example, reading it as an argument for perception as an *intervention on belief*, at Campbell and Cassam (2014: 79-80).

On Campbell's reading of this example, your conscious perception of the pig functions as an *intervention* on your beliefs.⁷ As already discussed, one of the critical features of an intervention, according to the recent causal literature, is that it be 'surgical', which means suspending all other influences on the effect in question. In this example, encountering the pig in the woods and perceiving it in good light (etc.) suspends any other influences there may be on your beliefs about pigs in the wood. You previously had doubts about there being a pig there, but the experience of perceiving this pig suspends that prior bias. It is not one consideration that you weigh against others: "Once it comes into play, it is the only factor specifically affecting what [you] believe on this point."⁸

⁷ Depending on your view of perception, you might think that what intervenes on your beliefs here is not the pig, but a conscious representation of the pig. Accordingly, you might think that there is a worry here that we can still doubt the veracity of the representation, and, hence, that the intervention is not strictly 'surgical', because it is being 'weighed' in the light of those other (skeptical) considerations. However, I take it that this is simply an instance of a *general* form of skepticism and as such poses no special threat to this view, since in all normal non-skeptical contexts, such conscious perception of the pig does in fact function in the way proposed here, having a *decisive impact* on your beliefs.

⁸ Campbell and Cassam (2014: 79).

It is interesting to compare this example with one involving a case of blindsight. Imagine how things would be different here if a blindsighted person was in the same situation: they had heard that there is a pig in these woods, but they think it is fairly unlikely. Now imagine that they encounter the pig in their blind field. They may *now* be inclined to guess that there is a pig there, if they were prompted, but this guess can—at best—only be weighed against their prior skeptical beliefs: it is simply one factor among many, and may indeed be outweighed by the strength of other factors.

Indeed, the typical blindsighted subject may not assign this evidence much weight at all, given that in actual cases of blindsight subjects always report that they have an experience of ‘guessing’ in such situations. That is, they experience their own judgment (when prompted) that there is a pig in their blind field as a judgment that is, phenomenologically speaking, no different to guesswork. And we do not typically assign judgments issuing from guesswork much weight in deliberation.

6.3 Conscious Perceptual Experience and ‘Zombie Action’

The examples just considered make it clear that consciousness—that is, conscious perceptual experience—regularly functions as a

surgical intervention on belief. Hence, we have a common example of the kind of intervention that was posited in the previous chapter: conscious perception typically does make a difference to agents' mental states. As I noted above, however, this is only part of the story of intentional realism: what is needed for the present argument is to *further* demonstrate that the very intentional states in question here are appropriately linked to behaviour — that they make a difference to behaviour. Without that, we don't have intentional realism.

What is needed is a view on which the conscious perceptual experience that informs an agent's mental states is also utilised in the control of *action*. Now, from one point of view, this claim might be seen as rather uncontroversial: there appears to be a fairly common pretheoretical assumption—even if the details are vague—that consciousness has some important role to play in what

happens when we act.⁹ And the involvement of *perceptual* consciousness seems to have a good level of intuitive support: it often appears that the way that we consciously perceive the world around us contributes importantly to our ability to act on that world.

From a different perspective, however, this commonsense view will be seen as seriously open to doubt. This is the perspective of a recent philosophical position that makes the case for so-called ‘Zombie Action’, on the basis of research in psychology and cognitive neuroscience.¹⁰ The Zombie Action hypothesis has been variously formulated, not always under that name, and can be briefly summarised as follows. The central claim is that our intentional actions are not controlled by conscious processes, but

⁹ For example, consider the essays in the recent collection *Does Consciousness Cause Behaviour?* (Pockett, Banks, and Gallagher 2006). They begin their introduction with the assertion that “All normal humans experience a kind of basic, on-the-ground certainty that we, our conscious selves, cause our own voluntary acts.” (1). Consider also, for example, the general reception of Freud’s theories of the unconscious sources of behaviour: the point is that they were generally regarded as *surprising* or somehow *contrary* to our normal self-understanding.

¹⁰ The term ‘Zombie action’ apparently comes from Koch and Crick’s (2001) article ‘The Zombie Within’.

are rather under the control of ‘Zombie systems’ — that is, sub-personal processes or systems that are remote from consciousness.

But what precisely is the ‘commonsense view’ of the relation between our conscious experience and action that is being disputed by Zombie Action? Above I said that the view is “that what we consciously experience contributes to the production of our bodily actions”, which remains at the intuitive level. We can sharpen the idea a little more by considering the following remarks from Brian O’Shaughnessy. Here O’Shaughnessy is considering the act of placing one’s index finger on the centre of a printed cross. In this case, he writes,

one keeps looking as one guides the finger, and does so right up until the moment the finger contacts the cross, and the reason, surely, is that sight is continually informing one as to where in one’s visual field to move one’s visible physical finger.¹¹

In this context, ‘sight’ is taken to refer to specifically *conscious* visual experience. I take it that this description gets something right about

¹¹ O’Shaughnessy (1992: 233).

how we commonly think about the relation between our conscious perception of the world, and our actions on that world. Conscious perception of an object *informs* our physical actions on that object. Further elaboration of this commonsense view is given by Andy Clark (albeit in the process of arguing against it). According to Clark, the commonsense view can be captured by the following definition:

Experience-Based Control (EBC). Conscious visual experience presents the world to the subject in a richly textured way; a way that presents fine detail (detail that may, perhaps, exceed our conceptual or propositional grasp) and that is, in virtue of this richness, especially apt for, and typically utilised in, the control and guidance of fine-tuned, real-world activity.¹²

On this view, EBC claims that conscious visual experience is utilised in “the control and guidance of fine-tuned, real-world activity.”

The ‘Zombie action’ arguments deny this connection between conscious perception and action: they dispute that

¹² Clark (2001: 4).

conscious visual experience is involved in the control of world-engaging activity in the way that commonsense (and EBC) supposes. Hence if the Zombie Action argument goes through, we have a problematic dissociation between the agent's mental states, and the actual control of action.

6.3.1 Perceptual demonstratives and the 'two-streams' view

How, then, is conscious perception involved in the guidance and control of such 'world-engaging activity', as claimed by EBC? It might be that, as John Campbell has suggested, perceptual demonstratives play an important role in the control of action, and that such perceptual demonstratives require specifically conscious perception of the demonstratively identified object. Hence the connection between consciousness and action control. But in what way are they involved?

The important step in Campbell's argument is to show that conscious attention is required for the proper understanding of perceptual demonstratives, on the grounds that it is necessary to 'single the object out' in one's conscious experience in order to understand which object is intended. A blindsighted person might have the object in their blind field, for example, and could thereby

utilise perceptual information regarding that object—perhaps in order to respond accurately in certain ‘forced choice’ situations involving the object—but this seems insufficient for the claim that they *know which* object is in question. (It is noteworthy on this point that blindsighted people do not ever seem to grow to ‘trust’ the accuracy of such guessing, and thereby adapt to rely on that perceptual information in a systematic way.)

Consider Campbell’s own example: suppose we are walking past a building with an array of windows and you say “that window is the window of my office”.¹³ I may be able to take in the whole building in a glance, and therefore know that the window to your office must be somewhere in my field of view. But that is not enough for me to know which window you mean: I must also single it out consciously, and *attend* to it.

Suppose you asked me to point to the window, but I have not yet managed to single it out consciously. I would claim that I could not point to the window, and if you asked me to guess I would also claim to be unable to guess. Now assume that you insist that I guess anyway, and that it turns out that I have the same capacity for better-than-chance accuracy at such tasks as do blindsighted

¹³ Campbell (2003: 151).

subjects. It still seems clear that I do not *know* which window you mean, and that, as in the case of blindsight, this is insufficient to understand the meaning of the demonstrative.

What then is the important connection between such (conscious) perceptual demonstratives and action control? The natural suggestion here is what he calls the *Grounding Thesis*.

Grounding Thesis. The meaning of a perceptual demonstrative is grounded in those aspects of perceptual experience that set the parameters for my action (how far I move, in what direction, and so on).¹⁴

According to the Grounding Thesis, the meaning that a perceptual demonstrative has on an occasion is fixed by the way in which I perceive the object: “that window” doesn’t mean the same thing on every occasion of use, but rather the meaning is given by information supplied by my perception of the object at that particular time. This ‘picture’ of the object provides information such as the size and shape of the object, and its location, or how far away it is from me, etc.

¹⁴ Campbell (2003: 152).

Perceptual demonstratives are centrally involved in action because it is this perceptually-grounded ‘picture’ of the object that sets the parameters for my action on the object: the information supplied in that picture is used to coordinate my action, i.e. how far I reach, and in what direction, and so forth. In this sense, it is consciously perceived information that is utilised for the control and guidance of real-world activity. Hence, we have the required integration between intentional states and action, both informed by our conscious perception of the world: indeed, this would apparently satisfy Clark’s characterisation of the commonsense view as well (EBC).

However, according to Campbell, the Grounding Thesis comes into conflict with certain empirical facts. In particular, the objection to this view, and to Experience-Based Control in general, comes from consideration of work done by David Milner and Melvyn Goodale.¹⁵ The argument is essentially another ‘Zombie Action’ problem for the Grounding Thesis: Campbell draws on the same set of empirical concerns as those which motivate the ‘Zombie

¹⁵ Their original monograph on the subject is Milner and Goodale (published 1995; second ed. 2006). See also the papers Milner and Goodale (1993) and the more recent and updated (2008).

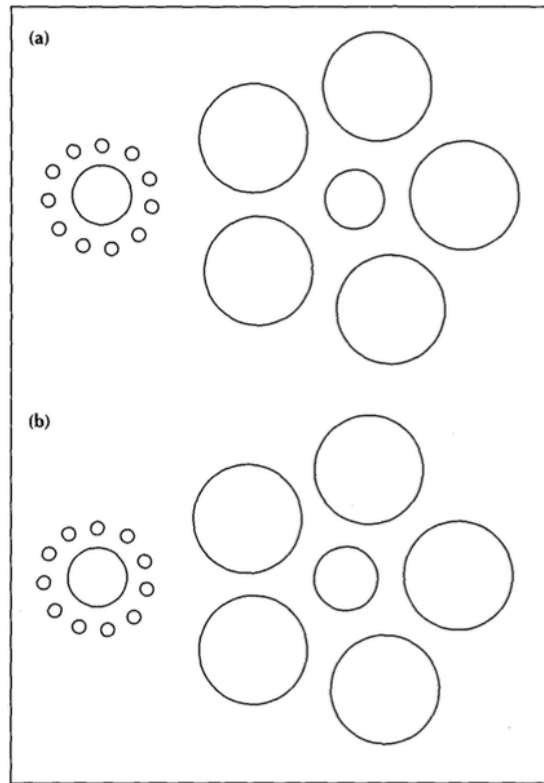
Action' literature, although he does not use that term. Nonetheless, the worry here is the same: that there would apparently be a large portion of human action that is controlled by processes or systems that are remote from consciousness.

According to the standard view of the primate visual system,¹⁶ there is an anatomical distinction between two visual streams that both stem from early cortical visual areas: the *dorsal stream* and the *ventral stream*. The work of Milner and Goodale suggests a view of the functional organisation of these two streams: their 'two systems' model of cortical visual processing makes a distinction between what they call 'vision for perception' and 'vision for action'. Roughly, the dorsal stream is characterised as being for direct motor control, and hence is referred to here as 'vision for action', while the ventral stream is for cognition, and is what is referred to as 'vision for perception'.

In general, philosophical attention has been focused on two main sources of evidence for the 'two systems' view, and it is this evidence that informs such Zombie arguments against the commonsense view of consciousness and action.

¹⁶ Mishkin and Ungerleider (1982).

The first source of evidence comes from consideration of pathological cases, in which patients have lesions to either the



ventral or dorsal stream. For example, Goodale and Milner report the case of DF, who suffered damage to the ‘perception’ pathway (ventral stream) resulting in visual agnosia: she was unable to recognise familiar objects or faces, or indicate the size, or orientation, of an object. Yet, when presented with an irregular object, she would grasp the object with the appropriate hand position, and when instructed to post a card through a slot, she was able to orient the card correctly and execute the movement without difficulty.

Conversely, for example, patients with optic ataxia resulting from lesions to the dorsal stream can accurately report on the size of the object (by holding up their fingers to show how big it is), or identify the correct orientation of a card. Yet, they are unable to properly grasp an irregular object, or to properly orient the card in order to post it through a slot.

The second source of evidence comes from experiments on normal adult subjects which reveal apparent perception-action disassociations, brought about when subjects are asked to act on objects featuring perceptual illusions, such as the Ebbinghaus Ring Illusion (*Fig. 6.1 a-b*).

*(Fig. 6.1 a-b) Ebbinghaus Ring Illusion*¹⁷

In the Ebbinghaus Ring Illusion (also called the ‘Titchener’ illusion), for most people, the inner circle can be made to appear either larger or smaller than its physical size, depending on the size of the surrounding annulus.

For example, in Figure 6.1a, the inner circles are in fact physically identical in size, but the circle on the left (surrounded by the small annulus) appears to be larger than the circle on the right.

¹⁷ Aglioti et al. (1995: 680).

By contrast, in Figure 6.1b the two inner circles appear to be the same size, but in fact the circle on the right (surrounded by the large annulus) is physically larger than the circle on the left.

In one experiment, Agliotti et al. used a version of the Ebbinghaus illusion in which the discs could be grasped (thin ‘poker chips’ were used as the circles). Subjects were asked to pick up one of the two target circles (the central, inner circles on the left or the right) depending on the *apparent size* of the circles.

Ultimately, the result was that even when both target circles appeared to the subject as though they were the same size, upon reaching for one of the discs, the grip aperture of thumb and forefinger was found to be calibrated to the *actual size* of the disc, and not the size that it appeared to be under illusion. Importantly, the size of the aperture was shown to have been set by *visual* information, yet it did not reflect the size of the object given in the subject’s conscious visual experience.

Return to Campbell’s first proposed explanation of the relation between perceptual demonstratives (which require conscious vision) and action. According to the Grounding Thesis, conscious experience does two things at the same time: it grounds the *meaning* of the demonstrative, and it also sets the *parameters for*

action on the identified object (via the informational ‘picture’ given in conscious experience). Campbell takes this empirical work to show that the Grounding Thesis cannot be right: it seems as though what is given in conscious experience is mediated by the ventral visual stream, while the parameters for motor control that coordinate the action on the object are actually set by information in the dorsal stream. As Clark puts it:

The kind of coding and processing implicated in the real-time guidance of delicate, fluent, object-engaged action is, it now seems, frequently and significantly distinct from that which supports our ongoing perceptual experience of the scene.¹⁸

Campbell’s move is to concede this point, and propose a modified version of the Grounding Thesis.

According to this modified view, instead of thinking that perceptual experience is required to set the parameters for action on the object (so-called ‘action programming’), we should think of that task as being informed by non-conscious information in the dorsal stream (as the Zombie argument claimed). However, this ‘action

¹⁸ Clark (2001: 17).

system' does not bind the various properties together, as happens in the perceptual system informing conscious vision; instead it supplies them 'one by one' for the control of action. Hence the properties that the object is represented as having—supplied by the 'Zombie' visual system—*individually* contribute to the motor configuration of the hand while reaching for the object.¹⁹

According to Campbell, this division creates a 'binding problem': when you demonstratively identify an object ("that is my hat") and decide to act on it, it is necessary for the action system to respond and configure your physical movements appropriately. It is therefore necessary to ensure that the perception and action systems are tracking one and the same object: that the information that is fed into the action system for motor control is information about the same object that is picked out in perception (e.g. "that hat").

Campbell's proposal is that the role of conscious perception is that of securing this initial connection. This is because conscious perception of the object will include information about the *location* of the object, and it is that information that is used to identify the object to the action system. That is,

¹⁹ Campbell (2003: 157).

conscious attention to the object will include some awareness of the location of the object, and [the] target for processing by the visuomotor system can be identified as: ‘the object at that location’²⁰

The claim is that location itself is the ‘binding principle’ that ensures the connection between perception and action: that the object you consciously intend to act on is in fact the object which is used to determine the coordinates of the action. Hence Campbell proposes the *Binding Thesis*:

Binding Thesis. Conscious attention is what defines the target of processing for the ‘action’ system, and thereby ensures that the object you intend to act on is the very same as the object with which the ‘action’ system becomes engaged.²¹

The analogy Campbell uses is that of a heat-seeking missile: it is necessary to point the missile in roughly the right direction, in

²⁰ Campbell (2003: 159).

²¹ *Ibid.* (160).

order for it to ‘lock on’ to the right object, but after that initial identification lower-level systems take over to track the object, and co-ordinate the real-time guidance of the missile. By picking out the object in consciousness—by identifying *that* thing as the one you mean to act on—it is possible for the system responsible for co-ordinating your physical movements to take over and track the relevant thing as ‘the object at *that* location’. Hence on this reading, perceptual consciousness has a direct role to play in action planning, although the real-time guidance of the fine motor activity involved in carrying out the physical movements is regulated by non-conscious systems.

How does Campbell’s considered view (the Binding Thesis) compare to the commonsense view noted above? First, consider the claim that is being made by the Zombie Action arguments. The best way to understand this is the way that Milner and Goodale put the point themselves:

Two Systems. The visual system that gives us our visual experience of the world is not the same system that guides our movements in the world.²²

This is one way of understanding what ‘Zombie Action’ theorists see as philosophically important in Milner and Goodale’s research. They make this claim on the basis of a specific reading of the two lines of evidence reported above: the pathological cases, and the perceptual illusions. And if *Two Systems* is true, then it seems like we have empirical evidence against the commonsense view.

Earlier, I claimed that the commonsense view might be sharpened by appealing to Clark’s thesis of ‘Experience-Based Control’:

Experience-based Control. Conscious visual experience ...[is] especially apt for, and typically utilised in, the control and guidance of fine-tuned, real-world activity.²³

²² Goodale and Milner (2004: 3). The label ‘two systems’ is mine: the subsequent definition is a quotation.

²³ Clark (2001: 4).

Now according to this reading of Milner and Goodale's empirical work, conscious visual experience typically *is not* utilised in the control and guidance of fine-tuned, real-world activity. Hence, it appears that according to this empirical work, we have shown the commonsense view—characterised here as EBC—to be false, and precluded the needed integration between conscious perceptual experience and action.

Campbell's suggestion is that conscious visual experience *does* have a role to play in action, but it is not the role that EBC says it is. Instead he claims, according to the Binding Thesis, that we should accept the general lesson of Milner and Goodale's research and concede that something like *Two Systems* is right. His novel suggestion here is that, although conscious visual experience might not play the role claimed for it by EBC, there is a *further* role not captured by that definition: that is, conscious visual experience allows us to demonstratively identify the object that we intend to act on, and enable the non-conscious 'Zombie' systems to lock on and modulate our fine-grained physical movements in the way that the research above suggests. So, this argument goes, conscious perceptual experience is required for action on an object after all,

since it is required to connect our intentions and our physical movements.

Does the Binding Thesis provide the required integration needed for intentional realism, and hence agency, as discussed above? I claim that it does not. In the previous chapter, intentional realism turned out to be a prerequisite for agency: it was suggested that such a realist view required the possibility of intervention on an agent's mental states. And in section 5.2 above, such interventions were shown to occur at the level of (e.g.) belief, as a result of our ability to consciously experience objects in the world. The variables that it is possible to intervene on here are those at the mental or intentional level.

But as I pointed out above, in order to complete this picture, we must show that the effects of the intervention do not stop there, as it were, and that the intentional states in question are appropriately connected to action. If we accept Campbell's Binding Thesis, we have not done enough to establish that connection.

According to the Binding Thesis, consciousness *only* gives us knowledge of reference—the object we intend to act upon is identified at *that* one—before turning over control to the non-conscious 'Zombie' systems that have been discussed above. Now it

is those systems which are responsible for providing the information that is used to guide action. Hence, it is those systems which *control* the action. So while consciousness gives us knowledge of reference, it is Zombie systems which give us knowledge *of the object*, and which thus control and co-ordinate the physical actions.

But the commonsense view has it that agents *consciously* control their actions — a view which was later unpacked more precisely as the claim that consciousness must be involved in the control and guidance of real-world activity (EBC). The Binding Thesis does not provide this control.

What is in fact needed is something closer to Campbell's original *Grounding* Thesis. According to this view, which Campbell had dismissed in response to the empirical concerns discussed above, it is conscious perceptual information that gives us knowledge of reference as well as providing the information about the object that is utilised in action control.

I take no stand on whether or not the Binding Thesis is true. In any case, it is not necessary to claim that the Binding Thesis is *false*, since Experience-Based Control (and the commonsense view in general) is compatible with the essential claim of the Binding

Thesis. Consciousness can still be involved in the selection-for-action that the Binding Thesis claims for it, but it must also be shown that, contrary to Zombie Action views, consciousness is directly involved in the real-time control and guidance of our actions.

6.3.2 Embodied demonstratives

Campbell dismissed the proposed Grounding Thesis because of the empirical evidence discussed above. But Christopher Mole disputes the philosophical interpretation of this evidence (the pathological cases, and the illusion studies).²⁴ He denies that the results of Milner and Goodale's work support the philosophical claims made on that basis, viz. the Zombie Action argument. The basis of Mole's claim is "that movement control and conscious experience are the work of *one and the same system*", which is a way of denying the *Two Systems* thesis, and hence undermining the philosophical interpretation of Milner and Goodale's research.²⁵

Now, as Wayne Wu has pointed out, there is a reading of Mole's claim here that makes it trivial: namely, that the visual

²⁴ Mole (2009; 2013).

²⁵ Mole (2009: 1002). Emphasis in original.

system as a whole yields both visual experience and visual guidance.²⁶ Making this claim non-trivial would require a criterion for system individuation, if the Zombie Action worry is not to turn into a simply verbal dispute. But the sameness or difference of systems is not really the salient point, as Mole acknowledges. The real issue—especially for present purposes—concerns the involvement of consciousness in action control.

In particular, Wu suggests that we should think of the issue as a matter of the consciousness or unconsciousness of the representations that control behaviour.²⁷ Now I take no stand here on whether we should adopt the representationalist approach, or whether Campbell's own relational view is to be preferred. The pertinent question here is whether, and to what extent, consciousness is directly involved in action control, as opposed to the view on which *non-conscious* systems or processes are primarily responsible for the control of action.

²⁶ Wu (2013: 219).

²⁷ Although as Mole points out, citing Bennett and Hacker, talk of *representations* being conscious or unconscious is better thought of in terms of a *subject's* being conscious of the content of a representation: "A representation of x is a conscious representation if and only if there is a subject who is conscious of x on account of x being the content of that representation." (Mole 2013: 232).

Mole answers this question affirmatively. He proposes that there is a class of *embodied demonstratives* that can contribute conscious information to the systems that control action. Hence, Mole's view is much closer to the Grounding Thesis that Campbell had originally proposed: the meaning of the embodied demonstrative is grounded in the information supplied by our conscious perception of the object, and which is thus used for action control.

To illustrate the point, return to the case of the visual form agnostic DF, who is unable to recognise the shape or orientation of a letterbox-like slot, although she is able to correctly co-ordinate her hand movements in order to post a card through that slot in the right way. According to some philosophers, this is evidence for the Zombie Action hypothesis: it suggests that the system responsible for her conscious visual experience of the slot—which is the system that is damaged—operates separately from the system that is controlling her action on the slot when she posts the card correctly. The more general 'Zombie' hypothesis is of course that, even when both systems are functioning normally, they are nonetheless operating in casually distinct ways. Mole sums up the worry about this 'mild form of epiphenomenalism':

it suggests that the causal chains leading to our experiences occur on branch lines from the causal chains that bring about our actions on the things that are experienced.²⁸

However, Mole suggests an alternate reading of the evidence on which Milner and Goodale's work does not license the Zombie Action conclusion. According to this view, the standard reading of Milner and Goodale's work explains the case of DF by claiming that her actions on the slot are guided by representations that make no contribution to her consciousness. (And that this is true in the normal, non-pathological cases as well.)

By contrast, Mole suggests an alternate interpretation of the evidence: according to this view, DF's action on the slot *is* guided by a representation that contributes to consciousness, "but that these representations have a format that is immediately useful only for informing D.F. of how the slot should be acted upon." What this means is that DF experiences the orientation of the slot "as the referent of a demonstrative concept that latches onto its content by

²⁸ Mole (2013: 233).

means of an embodied gesture.”²⁹ According to Mole’s view, when DF acts on the slot, by posting the card correctly, visual experience presents the slot to her, not as “vertical” or “at ninety degrees to the floor”, but as “being oriented so as to be acted on *thus*” — where ‘thus’ is characterised by a bodily gesture, the embodied demonstrative.

Because of this, Mole claims, DF’s action is not a ‘Zombie action’: it is “under the control of a conscious experience that has this embodied demonstrative character.”³⁰ It is on the basis of the conscious representation that she is able to act on the object in the way that she does, contrary to the Zombie Action hypothesis, because it is the conscious perception of the object that is providing the information necessary for action control, albeit not in a form that is available for verbal report, or other ways of reporting on the dimensions of the object. The information *is* available for action control, however, and is represented by the embodied demonstrative gesture.

Interestingly, one of the arguments Wu raises against Mole’s position takes the form of a *reductio*, based on the Ebbinghaus

²⁹ Mole (2013: 235).

³⁰ Mole (2013: 235).

reaching illusion. According to this argument, stated informally, if we deny that Milner and Goodale's Two Systems hypothesis (or some similar view) is the right way of explaining the results of the reaching experiment, we seem to end up in a difficult position.

First, note that there is a visual representation controlling the action (the reaching behaviour) that represents the disc as size x . Second, there is a (conscious) visual representation informing the verbal report that represents the disc as size y , and where $y \neq x$. If we assume, for reductio, that the representation x is conscious (the representation informing the reaching behaviour), then we seem to be in a situation where the subject has a conscious visual representation of the disc as size x and size y — that is, we end up saying that the disc *looks* to the subject to be both size x and y . But, the argument goes, the disc *cannot* look that way, hence the representation x is not conscious.

The nature of the reaching illusion is such that according to the subject's verbal report, which everyone agrees is informed by the subject's conscious visual experience, the disc *looks* a certain size (y). But because of the effects of the illusion, this is not the actual, physical size of the disc. Yet when the subject reaches for the disc, the grip aperture *is* set to the actual size of the disc (x), meaning

that the reaching must be guided by some other representation. Now, Wu believes that if the subject were conscious of this representation of the disc's size (x), this would mean that the disc *looked* that way to the subject too. Hence, the claimed absurdity: the disc looks to the subject to be size x *and* size y .

Mole's account of embodied demonstratives offers a response to this argument. According to Mole's view, the reductio does not follow from those premises: the claim that leads to the absurdity ("but the disc *cannot* look that way . . .") is an *assumption* introduced by Wu. In fact, Mole denies this claim. Since, on his view, x and y are represented under different modes of presentation, it may not be manifest to the subject that the sizes are not the same — i.e. the size represented for verbal report (y), and the size represented for action guidance (x).³¹ Hence it is open to Mole to accept the apparently contradictory claim that the subject 'represents' the disc as being both 'size x ' and 'size y '.

6.4 Disputing the Evidence for Zombie Action

Mole has much more to say in defence of the general plausibility of the 'embodied demonstrative', but it would take us too far away

³¹ Mole (2013: 237).

from the central point to discuss the idea in more detail. What is interesting to note here is that it represents an advancement over Campbell's use of the perceptual demonstrative in his own Binding Thesis, since according to Mole the embodied demonstrative is a means for conscious information to be utilised in the guidance and control of action — as opposed to the more limited role of consciousness allowed by Campbell's own view.

In any case, what I would like to focus on here is the exchange between Mole and Wu that was just considered, in particular regarding the putative 'absurdity' involved in apparently consciously perceiving contradictory content (and which serves as a core premise in the Zombie Action argument). Mole's response was built into his notion of the embodied demonstrative. On that view, the differing size-contents of the two representations of the discs are actually represented under *different* modes of presentation, and hence the contradiction disappears.

In this section, I will complete the argument against Zombie Action by generalising this strategy, and criticising the Zombie arguments in a way that does not depend on any particular philosophical theory (such as Mole's embodied demonstratives).

Firstly, it is worth pausing to consider in more detail what does and does not need to be established by the defender of EBC. Importantly, it does not need to be shown that *no* sub-personal, non-conscious systems or processes are involved in the production of action. Human beings are physical, embodied agents, and it would be highly implausible to suppose that the aetiology of action involved no causal processes that were non-conscious, including non-conscious mental processes. It is no barrier to accepting the causal relevance of conscious perceptual information on action control that it is not the *sole* causal factor controlling action.

With that in mind, it will have to be quite specific empirical evidence that can support the Zombie Action hypothesis. For it to be philosophically interesting, and actually come in to conflict with EBC, it would have to be shown that conscious perceptual information clearly is *not* playing such a casual role — for example, by showing that non-conscious processes alone are *causally sufficient* for the control of action. Simply demonstrating that certain features of our physical actions are modulated by processes of which we are not (and perhaps could not be) conscious is not sufficient to repudiate EBC in favour of Zombie Action.

The strategy in Mole's paper was to dispute the interpretation of Milner and Goodale's work on visual perception that leads to the Zombie Action conclusion. In particular, he disputes the interpretation of the experiment involving DF's action on the letterbox-like slot. Proponents of the Zombie Action hypothesis take this experiment to reveal a disassociation between conscious perception and action control: since DF is unable to report on the orientation of the slot, but is able to reliably post a card through the slot, this is taken to show that the action is controlled by non-conscious processes. But Mole's suggestion is that we could also take this to show that DF only has access to the information about orientation that is revealed in conscious perception when it is positioned as a target for action, and is not available for report other than in the form of the embodied demonstrative gesture.

At best, the evidence can be seen as equivocal between these competing interpretations of the data: support for Zombie control, on the one hand, and support for embodied demonstratives, on the other. If one is not independently convinced about the general theory of embodied demonstratives, Mole's response might not be seen as decisive. However, the general strategy of Mole's response

here is instructive: it is worth looking more carefully at what pathological cases such as DF, and the research on perceptual illusions, really licenses us to say.

First, return to the Ebbinghaus ring illusion. In this case, discussed above, it was claimed that there is a dissociation between the conscious perception of the disc, and action on the disc. The ‘Zombie’ interpretation is that, since subjects’ reports on the size of the disc reveal the effects of the illusion, and since their physical actions seem to be informed by the real size of the disc, the latter cannot be controlled by conscious perception — that is, it cannot be informed by information that comes from the subject’s conscious perceptual experience.

But why should we suppose this? Morgan Wallhagen has argued, in response to a different experiment involving the illusory movement of a target, that such cases depend on the assumption that visual experience cannot represent veridical and illusory contents. This is a general version of Mole’s response, but without the more specific claim that the illusory and veridical contents must be represented under different modes of presentation, and is not dependent on the theory of embodied demonstratives given by Mole. Nonetheless, it adopts the same strategy: effectively

challenging the premise in the Zombie Action argument which assumes that veridical contents cannot accompany the illusory contents in conscious perception. According to Wallhagen's view, there are actually many cases in which both are presented in conscious visual experience.

For example, consider the Müller-Lyer illusion. In this case, subjects will report having an experience as of the two lines being an unequal length, whereas the physical dimensions of the lines are equal. Wallhagen suggests a simple way to demonstrate that veridical contents are also presented in conscious perception: draw the lines on graph paper. One can then simply observe that the ends of the Müller-Lyer lines are touching parallel vertical graph lines.

If it were the case that the content of perceptual experience was *exhausted* by the illusory contents, it would not be possible to observe that the lines are of equal length in this way. The content of visual experience here contains veridical *and* illusory "length contents".³² In normal cases, the subject simply doesn't notice that their visual states have this content. Similar examples given by Wallhagen include Escher drawings: in these cases, there is what he calls the "objective content", which is the impossible properties that

³² Wallhagen (2007: 554).

no objective object could have; yet there is also the “line content”, which is perfectly coherent — “there are just lines of certain lengths on paper”.³³

That there are such further examples of illusory and veridical contents being presented in conscious visual experience, and since there is nothing in the description of the Ebbinghaus illusion experiment conducted by Milner and Goodale that rules out such an interpretation, I suggest that this is a plausible response to such results — that the reaching behaviour is dependent on veridical information that is presented in conscious visual experience.

Further, if one is convinced by Wallhagen’s examples here (e.g. the Müller-Lyer lines and the Escher drawings), then it might not even be necessary to claim that the illusory and veridical contents must be represented under different modes of presentation. However, since that claim is itself contentious, I will not directly argue for it here: the conclusion established by Mole’s argument considered earlier is also sufficient to make the point. Whether they are presented under different modes of presentation or not, the claim here is that veridical and non-veridical contents can be presented in conscious visual experience.

³³ Wallhagen (2007: 554, n. 11).

In illusion experiments of this kind, the subjects are typically assumed to have normal perceptual capacities. But in the case of DF, frequently discussed by proponents of the Zombie Action argument, we are dealing with a pathological case in which part of the normal perceptual system is not functioning. It might be thought that the Zombie Action argument is stronger here because the subject is (so it is claimed) physiologically incapable of consciously perceiving shape and orientation. Hence, there is (apparently) no way that conscious perception of the physical dimensions of the letterbox could be informing DF's actions on the object, because there *is no* such perception of those features of the object.

However, we should be very careful when characterising what it is that DF is supposedly unable to do, before concluding that the Zombie Action view offers the best interpretation of the results of her trials in the experimental situations. DF has the condition known as visual form agnosia, which means that she performs very poorly at tasks designed to test her sensitivity to the physical form of objects (e.g. shape, orientation, etc.). In much of the literature, however, there is very little information about the phenomenology of her perceptual experience. Indeed, it is

strikingly hard to imagine what her experience of objects must be like, given that she apparently retains the capacity to perceive properties such as texture and colour, while being apparently unable to discern shape or the physical dimensions of the object.

On the contrary, it would seem that there ought to be *some* basic sense in which she is able to perceive form, viz. as the boundary or limit of the coloured, textured areas that we are told she is able to perceive. Similarly, we are told that perimetry tests indicate that her visual field is not, for example, limited to very small regions of space (i.e. such that she is not able to take in the boundaries of such objects at all). Certainly, we are not given any better explanations of her phenomenology: in the absence of that, we should assume that there is *some* sense in which she is able to consciously perceive form. So what is going on in the experimental situations?

What we do know about DF's situation is that she is unable to *report* on the form of the object: she cannot *answer questions* about the form of the object, or *indicate* the orientation of the object using her hands, etc. Note that the claim that she is 'unable to consciously perceive form' is a much stronger claim than any of these, and is not licensed or entailed by any of them.

The fact that, despite her visual agnosia, she is able to act on the object in a way that would require information about physical form is taken by proponents of the Zombie argument to be evidence that non-conscious ‘Zombie systems’ are controlling the action: but this is only a legitimate move if one *assumes* the stronger view of her condition just noted. Without this assumption, which is simply quite implausible, her ability to act on the object is perfectly consistent with the commonsense view according to which her conscious perception includes some kind of information about form—but which is, for whatever pathological reason, *unavailable for report*—and which is being used to control the action. Indeed, in the absence of more positive evidence that she clearly cannot have conscious experience of form in any sense, this reading of DF’s situation is the *more* plausible one.

Interestingly, in his own positive account of DF’s condition, Wallhagen suggests a view according to which her problem is one of knowledge, not consciousness. He suggests that what is problematic about her situation is that she is unable to bring the objects of her experience ‘under concepts’, and it is this fact that explains her inability to report on the form of objects. In addition, he suggests that we appeal to *nonconceptual* content to explain how

DF nonetheless *experiences* the form of objects, while at the same time lacking the *conceptual* abilities just noted (as a result of her visual agnosia).

While I think it is important to look carefully at DF's condition before taking it as supporting evidence for the Zombie Action view, the details of Wallhagen's alternative proposal are less convincing. The move seems to be too quick: i.e. that the philosophical claim that DF's condition should be interpreted as an inability to bring objects of experience under concepts is the best way of characterising the effects of that specific damage to the ventral stream. We would need to know much more about the relation between physical information about the brain and the psychological phenomena that we are interested in here before such a conclusion could be plausibly entertained. Wallhagen's suggestion raises more philosophically contentious questions than it answers. In any case, the salient point here is not the positive proposal about what *is* going on in DF's situation, but rather what *is not* happening — that is, the extent to which experimental research on her condition fails to support the Zombie Action argument.

The problem of this chapter was to show that consciousness makes a difference to behaviour. I conclude that we have good reason to believe that it does. The main challenge to this largely commonsense view was the Zombie Action argument, or various forms of it. By contrast, the arguments of this chapter have shown that we need not accept the conclusions that the Zombie arguments seem to force upon us.

7

Conclusion

7.1 The Problem Reconsidered

I said in the Introduction that this project has two central arguments, one ‘first-order’ and one ‘second-order’. I addressed the second-order argument first, concerning the shape of the free will problem. I call it a second-order argument because it is really an argument about certain existing arguments, i.e. those making up the ‘free will problem’.

The main thrust of my argument in this regard has been that we should stop hanging on to the terminology of ‘free will’, and associated terms, and stop allowing it to shape our ongoing perspective of the philosophical problems. My main suggestion was to just look at what philosophers substantially disagree over when they disagree about free will: most of the time it is really *moral responsibility* that is the problem. My practical suggestion was simply to define the term operationally as whatever those

conditions are that an agent must meet in order to be morally responsible. Then we have a well defined use for the term, and can continue to use it. Of course, we don't *need* to retain the term at all, given my view of things. Whether we should in fact retain the term at all is not of direct philosophical concern to me here, but those are the options as I see them.

The interesting part comes when we realise what this conceptual housekeeping has allowed us to see more clearly: that there are certain things in the existing literature that aren't about moral responsibility at all, but are about *agency*. So here is an important *distinction* that was obscured, and that my perspective helps bring into view.

Hence the *first-order* part of the project is about this concept of agency, and how it turns out to play a role in what has come to be called 'the free will problem'. Of course, *agency* is not a concept that is unique to this problem. It has its own burgeoning literature, where it is addressed on its own terms. For this very reason, what I have said about agency here clearly does not amount to a full-scale, novel account of human agency. That would take at least a book on its own: the fact is that I have to strike a dialectical balance between

actually investigating agency itself, and situating this investigation in the context of a larger *conceptual* claim.

So although I do make direct metaphysical claims about agency in a non-trivial way, the added disclaimer is that this is of necessity only a *sketch* of a theory of agency. The point has been to show how there is in fact a plausible account of agency that is consistent with the role I identified for it within the free will dialectic.

Looking at it from another perspective, I am claiming that the structure of agency itself turns out to be directly relevant to arguments about free will. It is not therefore possible to investigate such ‘free will’ independently of the direct study of *agency*.

7.2 Causal Integration as Central to Agency

I have shown how agency is relevant to the ‘free will and moral responsibility’ debate. I gave two primary examples of cases where the theory of agency is doing the ‘heavy lifting’ in certain arguments that are putatively about ‘free will’.

A central notion that came out of these examples was that of ‘causal integration’. I claimed that causal integration is a central feature of agency, and showed how it played a role in responding to

Pereboom's Four Case argument, for example. So the notion of agency as causal integration was central to an argument against Pereboom's hard incompatibilism.

The further challenge (now the first-order part of my argument) was to show that this kind of causal integration actually is part of a plausible theory of agency. To that end, I sketched the outlines of such a theory of agency. The notion of causal integration was unpacked further, and I suggested that—at least in the human case—agency is a matter of the causal integration of phenomenal consciousness, the agent's intentional states (such as belief and desire), and behaviour. So, for example, what goes wrong in Pereboom's 4CA is that the agent's intentional states are rendered epiphenomenal, thus breaking the causal integration between them and the agent's behaviour. Hence we do not have an agent at all in this case, contrary to Pereboom's own premise. So the argument fails to go through, but not because of anything about 'free will': it fails because it makes assumptions about agency that do not stand up to scrutiny.

The notion of causal integration, then, is my central substantive claim about agency. In the first instance, when discussing the 4CA, I agreed with Demetriou's argument when she

pointed out that there is something defective about Plum's agency in that case. She had claimed that he was not a 'causally integrated entity', and illustrated this with the schematic diagram that was reproduced above. The point was that Plum's mental states were divorced from the causation of his behaviour in a way that was deeply at odds with both our commonsense view of agency, and most philosophical accounts. Plum's mental states were epiphenomenal.

What was not said in Demetriou's argument was what the needed causal integration should look like. This is not a criticism of the argument, given the context of her discussion of Pereboom's work. What it does mean is that the rather abstract claim that Plum fails to be a 'causally integrated entity' remains compatible with many *positive* views of what a causally integrated entity should look like. All that is really suggested is what's implied by negation: we are to assume that an agent's mental states *should* be causally connected to his physical behaviour, in a way that Plum's are not.

I have claimed that an agent's mental states should be causally connected to her behaviour in this way, and that these mental states should be intentional, and roughly as characterised by our folk psychology — in short, that *intentional realism* is true. For

that reason, I made the case for accepting a form of intentional realism, against the challenges from eliminativism and instrumentalism.

Strictly speaking, the truth of intentional realism is by itself sufficient to satisfy the ‘causal integration’ criterion that I have just discussed. For two reasons, however, I did not leave the discussion there. Firstly, as part of the discussion in Chapter 4, I claimed that a plausible account of agency also contains a notion of *subjectivity* that is not already captured by the basic structure of intentional realism. This was inspired by Helen Steward’s suggestive remark in her discussion of the topic, that it is part of our commonsense or intuitive conception that an *agent* is the ‘centre of some form of subjectivity’.

Secondly, and more importantly, I claimed that—as a matter of fact—we need to bring in the idea of consciousness in order to properly make the case for intentional realism. So although, in principle, the claim of intentional realism is by itself sufficient for causal integration, in order to actually *get* intentional realism we need to appreciate the role of phenomenal consciousness. And, of course, we have independent reason to make consciousness a necessary part of agency because of the *subjectivity* claim.

The picture of agency that I have sketched here is one that is built around the notion of causal integration. The two central examples I cited above (the 4CA, and ‘agent causation’) show how this notion makes a difference to arguments putatively about ‘free will’. I have elaborated on the notion of causal integration to show that there is a *possible* account of agency that takes that notion as its central feature. But I have also gone some way to showing that this account of agency is a *plausible* and maybe *true* account of human agency. Even if it is, I have not claimed that it’s a *complete* account of human agency (not to mention non-human agency). In the end, the main thrust of my argument here has been that we cannot begin to address the ‘free will problem’, such as it is, without first considering our understanding of agency: at the very least, the two are not nearly as independent as has been previously thought.

Bibliography

AGLIOTI, S., DESOUZA, J., and GOODALE, M. 1995. 'Size-contrast illusions deceive the eye but not the hand', *Current Biology*. 5(6): 679-85.

ALLEN, S. 2012. 'What matters in (naturalised) metaphysics?', in *Essays in Philosophy* (13).

ALVAREZ, M. and HYMAN, J. 1998. 'Agents and their actions', *Philosophy*. 73(2): 219-45.

ANSCOMBE, G. E. M. 1959. *An Introduction to Wittgenstein's Tractatus*. Thoemmes Press.

——— 1963. *Intention*. (Second Edition). Cambridge, MA: Harvard University Press.

——— 1971. 'Causality and determinism', in ANSCOMBE, G. E. M. (ed.) *Metaphysics and Philosophy of Mind, Collected Papers, Vol. II*. Oxford: Blackwell.

AUSTIN, J. L. 1962. *How to Do Things with Words*. (Second Edition).
SBISÀ, M. and URMSON, J. O. (eds.) Oxford: Oxford University Press.

AYER, A. J. [1954]. 2002 'Freedom and Necessity', in *Philosophy of Mind: Classical and Contemporary Readings*. CHALMERS, D. (ed.): 662-6. Oxford: Oxford University Press.

BALAGUER, M. 2010. *Free Will as an Open Scientific Problem*. London: MIT Press.

BARGH, J. 2008. 'Free will is un-natural', in BAER, KAUFMANN, and BAUMEISTER (eds.) *Are We Free? Psychology and Free Will*. 128-154.

——— and FERGUSON, M. 2000. 'Beyond Behaviourism: On the automaticity of higher mental processes, *Psychological Bulletin*. 126: 925-45.

BAUMGARTNER, M. 2013. 'Rendering Interventionism and Non-Reductive Physicalism Compatible', *Dialectica*. 67(1): 1-27.

BAYNE, T. 2008. 'The Phenomenology of Agency', *Philosophy Compass*. 3(1): 182-202.

BENNETT, K. 2003. 'Why the Exclusion Problem Seems Intractable, and How, Just Maybe, to Tract It', *Noûs*. 37(3): 471-97.

——— 2007. 'Mental Causation', *Philosophy Compass*. 2(2): 316-37.

BERMÚDEZ, J. L. 2005. *Philosophy of Psychology*. Routledge.

BISHOP, R. 2006. 'The Hidden Premise in the Causal Argument for Physicalism', *Analysis*. 66: 44-52.

BLOCK, N. 1989. 'Can the mind change the world?', in BOOLOS (ed.) *Meaning and Method: Essays in Honour of Hilary Putnam*. Cambridge: Cambridge University Press.

——— 2003. 'Do Causal Powers Drain Away?', *Philosophy and Phenomenological Research*. 47 (1): 133-50.

BOTTERELL, A. 'Review: *A Physicalist Manifesto: Thoroughly Modern Materialism* by Andrew Melnyk', *The Philosophical Review*. 114(1): 125-28.

BRATMAN, M. 2000. 'Reflection, Planning, and Temporally Extended Agency', *The Philosophical Review*. 109(1): 35-61.

BROWN, R. and LADYMAN, J. 2009. 'Physicalism, Supervenience, and the Fundamental Level', *The Philosophical Quarterly*. 59(234): 20-38.

CAMPBELL, J. 2003. 'The Role of Demonstratives in Action Explanation', in ROESSLER, J. and EILAN, N. (eds.) 150-64. *Agency and Self-Awareness: Issues in Philosophy and Psychology*. Clarendon Press.

——— 2007. 'An Interventionist Approach to Causation in Psychology', in GOPNIK, A. and SCHULZ, J. (eds.) *Causal Learning: Psychology, Philosophy and Computation*. 58-66. Oxford: Oxford University Press.

——— 2010. 'Independence of Variables in Mental Causation', *Philosophical Issues*. 20: 64-79.

——— 2010b. 'Control Variables and Mental Causation', *Proceedings of the Aristotelian Society*. 110(1): 15-30.

——— and CASSAM, Q. 2014. *Berkeley's Puzzle: What Does Experience Teach Us?* Oxford: Oxford University Press.

CAMPBELL, J. K. 2011. *Free Will*. Polity Press.

CHALMERS, D. 1996. *The Conscious Mind*. New York: Oxford University Press

——— 1999. 'Materialism and the Metaphysics of Modality', *Philosophy and Phenomenological Research*. 59: 473–493.

——— 2011. 'Verbal Disputes', *Philosophical Review*. 120(4): 515-566.

CHILD, W. 1996. *Causality, Interpretation, and the Mind*. Oxford: Clarendon Press.

CHISHOLM, R. 1976. *Person and Object: A Metaphysical Study*. Illinois: Open Court Publishing Company.

——— 1978a. 'Comments and Replies', *Philosophia*. 7(3-4): 597-636.

——— 1978b. 'Is there a mind-body problem?', *Philosophical Exchange*. 2: 25-34.

CHOMSKY, N. 1972. *Language and Mind*. New York: Harcourt Brace Jovanovich.

——— 1995. 'Language and Nature', *Mind*. 104(413): 1–61.

——— 1998. 'Comments: Galen Strawson, *Mental Reality*', *Philosophy and Phenomenological Research*. 58(2).

CHURCHLAND, P. 1981. 'Eliminative Materialism and the Propositional Attitudes', *Journal of Philosophy*. 78(2): 67-90.

CLARK, A. 2001. 'Visual Experience and Motor Action: Are the Bonds Too Tight?', *Philosophical Review*. 110(4): 495-519.

CLARKE, R. 1995. 'Indeterminism and Control', *American Philosophical Quarterly*. 32(2): 125-38.

——— 1996. 'Contrastive rational explanation of free choice', *The Philosophical Quarterly*. 46: 185-201.

——— 2003. *Libertarian Accounts of Free Will*. Oxford: Oxford University Press.

——— 2009. 'Dispositions, Abilities to Act, and Free Will: The New Dispositionalism', *Mind*. 118: 323-51.

——— 2010. 'Because She Wanted To', *The Journal of Ethics*. 14(1): 27-35.

——— 2014. 'Agency and Incompatibilism', *Res Philosophica*. 91(3): 519-25.

CORNFORD, F. M. 1979. *Plato's Theory of Knowledge*. London: Routledge and Kegan Paul.

CRANE, T. 1991. 'Why Indeed? Papineau on Supervenience', *Analysis*. 51(1): 32-7.

——— and MELLOR, D.H. 1990. 'There is no Question of Physicalism', *Mind*. 99: 185-206.

——— and ÁRNADÓTTIR, S. T. 2013. 'There is no Exclusion Problem', in GIBB, S. C., LOWE, E. J., INGTHORSSON, R. D. (eds.) 2013. *Mental Causation and Ontology*, pp. 248-66. Oxford: Oxford University Press.

CRAVER, C. 2007. *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Oxford: Oxford University Press.

CROOK, S. and GILLET, C. 2001. 'Why physics alone cannot define the "physical": materialism, metaphysics, and the formulation of physicalism', *Canadian Journal of Philosophy*. 31(3): 333-60.

DAHLBOM, B. (ed.) 1993. *Dennett and His Critics*. Oxford: Basil Blackwell.

DAVIDSON, D. 1963. 'Actions, Reasons, and Causes', *The Journal of Philosophy*. 60(23): 685-700.

——— 1973-4. 'On the Very Idea of a Conceptual Scheme', *Proceedings of the American Philosophical Association*. 47: 5-20.

——— 1973. 'Freedom to Act', in *Essays on Freedom of Action*. Honderich, T. (ed.): 137-56. London: Routledge & Kegan Paul.

——— 1980. *Essays on Actions and Events*. Oxford: Oxford University Press.

——— 1986. 'A Coherence Theory of Truth and Knowledge', *Truth and Interpretation: Perspectives on the Philosophy of Donald Davidson*. LEPORE, E. (ed.) Oxford: Blackwell.

DEMETRIOU, K. 2010. 'The Soft-Line Solution to Pereboom's Four-Case Argument', *Australasian Journal of Philosophy*. 88(4): 595-617.

DENNETT, D. C. 1973. 'Mechanism and Responsibility', in Honderich, T. (ed.) *Essays on Freedom of Action*. London: Routledge & Kegan Paul.

——— 1981. *Brainstorms: Philosophical Essays on Mind and Psychology*. MIT Press.

——— 1986. *Content and Consciousness*. (2nd Ed.) Routledge and Kegan Paul.

——— 1988. *The Intentional Stance*. Cambridge, MA: The MIT Press.

- 1988b. 'Précis of *The Intentional Stance*', *Behavioural and Brain Sciences*. 11: 495-546.
- 1991. 'Real Patterns', *Journal of Philosophy*. 88: 27-51.
- 1992. *Consciousness Explained*. Back Bay Books.
- 1993. 'Back from the Drawing Board', in DAHLBOM (ed.) *Dennett and His Critics*, pp. 201-35. Oxford: Basil Blackwell.
- 1996. *Darwin's Dangerous Idea*. Simon and Schuster.
- 1997. *Kinds of Minds: Towards and Understanding of Consciousness*. Basic Books.
- 2000. 'With a Little Help From My Friends', in ROSS (et al. ed.) *Dennett's Philosophy: A Comprehensive Assessment*, pp. 327-88.
- 2005. *Sweet Dreams: Philosophical Obstacles to a Science of Consciousness*. Bradford Books.
- DOWE, P. 2000. *Physical Causation*. New York: Cambridge University Press.
- ELSTER, J. 2007. *Explaining Social Behaviour: More Nuts and Bolts for the Social Sciences*. Cambridge: Cambridge University Press.
- FELDMAN, F. and Skow, B. 2015. 'Desert', *Stanford Encyclopaedia of Philosophy*. Winter 2015. Zalta, E. (ed.) [<http://plato.stanford.edu/archives/win2015/entries/desert/>]
- FISCHER, J. M. (ed.) 1989. *God, Foreknowledge, and Freedom*. Stanford University Press.
- 1994. *The Metaphysics of Free Will: An Essay on Control*. Blackwell.
- 2006. *My Way: Essays on Moral Responsibility*. Oxford University Press.
- 2010. 'The Frankfurt Cases: The Moral of the Stories', *Philosophical Review*. 119(3): 315-36.

FISCHER, J. M. and RAVIZZA, M. 1994. *Perspectives on Moral Responsibility*. Cornell University Press.

——— 1998. *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge University Press.

FODOR, J. 1974. 'Special Sciences: Or the Disunity of Science as a Working Hypothesis', *Synthese*. 28: 97–115.

——— 1975. *The Language of Thought*. New York: Thomas Crowell.

——— 1987. *Psychosemantics*. Cambridge, MA: The MIT Press.

——— 1994. *A Theory of Content and Other Essays*. Cambridge. MIT Press.

——— 1997. 'Special Sciences: Still Autonomous After All These Years', in TOMBERLIN, J. (ed.) *Philosophical Perspectives 11: Mind, Causation, and World*, pp. 149-64. Boston: Blackwell.

FOOT, P. 1967. 'The Problem of Abortion and the Doctrine of the Double Effect', *Oxford Review*. No. 5.

FRANKFURT, H. 1969. 'Alternate Possibilities and Moral Responsibility', *The Journal of Philosophy*. 66(23): 829-39.

FRANKLIN, C. E. 2011. 'Farewell to the luck (and *Mind*) argument', *Philosophical Studies*. 156: 199-230.

——— 2011b. 'The Problem of Enhanced Control', *Australasian Journal of Philosophy*. 89(4): 687-706.

——— 2015. 'Everyone thinks that an ability to do otherwise is necessary for free will and moral responsibility', *Philosophical Studies*. 172(8): 2091-107.

——— forthcoming. 'If Anyone Should Be an Agent-Causalist, then Everyone Should Be an Agent-Causalist'

GAZZANIGA, M. 2011. *Who's in Charge? Free Will and the Science of the Brain*. Harper Collins.

GIBB, S. C., LOWE, E. J., INGTHORSSON, R. D. (eds.) 2013. *Mental Causation and Ontology*. Oxford: Oxford University Press.

GINET, C. 2007. 'An Action Can Be Both Uncaused and Up to the Agent', in LUMER, C. and NANNINI, S. (eds.) *Intentionality, Deliberation and Autonomy: The Action-Theoretic Basis of Practical Philosophy*. 243–55. Ashgate.

GLYMOUR, C. 1999. 'A Mind is a Terrible Thing to Waste', *Philosophy of Science*. 66(3): 455-71.

——— 2001. *Bayes Nets and Graphical Causal Models in Psychology*. MIT Press.

GOFF, P. 2010. 'Ghosts and Sparse Properties: Why Physicalists have More to Fear from Ghosts than Zombies', *Philosophy and Phenomenological Research*. 81(1): 119-39.

GOODALE, M. and MILNER, D. 1992. 'Separate visual pathways for perception and action', *Trends in Neurosciences*. 15(1): 20–5.

——— 2008. 'Two visual systems re-viewed', *Neuropsychologica*. 46(3): 774-85.

GREENE, J. 2008. 'The Secret Joke of Kant's Soul', in SINNOTT-ARMSTRONG, W. (ed.) *Moral Psychology (Vol. 3)*: 35-80. MIT Press.

HEMPEL, C. G. 1969. 'Reduction: Ontological and Linguistic Facets', in MORGENBESSER and SUPPES (eds.) *Philosophy, Science, and Method: Essays in Honour of Ernest Nagel*, pp. 179–99. New York: St. Martin's.

HILL, C. And MCCLAUGHLIN, B. 1999. 'There Are Fewer Things in Reality Than Are Dreamt of in Chalmers's Philosophy', *Philosophy and Phenomenological Research*. LIX(2): 445-54.

HITCHCOCK, C. 1996. 'The Role of Contrast in Causal and Explanatory Claims', *Synthese*. 107: 395–419.

——— 2009. 'Structural equations and causation: six counterexamples', *Philosophical Studies*. 144(3): 391-401.

HOERL, C., MCCORMACK, T. and BECK, S. (eds.) 2011. *Understanding Counterfactuals, Understanding Causation: Issues in Philosophy and Psychology*. Oxford: Oxford University Press.

HOHWY, J. and KALLESTRUP, J. (eds.) 2008. *Being Reduced: New Essays on Reduction and Explanation in the Special Sciences*. Oxford: Oxford University Press.

HONDERICH, T. 1988. *A Theory of Determinism: The Mind, Neuroscience and Life Hopes*. Clarendon Press.

HORGAN, T. 1989. 'Mental Quasation', *Philosophical Perspectives*. 3: 47-76.

——— 2001. 'Causal compatibilism and the exclusion problem', *Theoria*. 16: 95-116.

——— 2007. 'Mental Causation and the Agent-Exclusion Problem', *Erkenn*. 67: 183-200.

——— forthcoming. 'Causal compatibilism about agentive phenomenology', in HORGAN, T., SABATES, M., and SOSA, D. (eds.) *Supervenience in mind: essays in honour of Jaegwon Kim*. Cambridge: MIT Press.

——— and TIMMONS, M. 2011. 'Introspection and the phenomenology of free will: problems and prospects', *Journal of Consciousness Studies*. 18(1): 180-205.

HORGAN, T., TIENSON, J. and GRAHAM, G. 2003. 'The Phenomenology of First-Person Agency', in WALTER, S. and HECKMANN, H-D. (eds.) *Physicalism and Mental Causation*. Imprint Academic.

HORNSBY, J. 2001. *Simple Mindedness: In Defence of Naïve Naturalism in the Philosophy of Mind*. Harvard: Harvard University Press.

——— 2004. 'Agency and Alienation', in DE CARO, M. and MACARTHUR, D. (eds.) *Naturalism in Question*, pp. 173-87. Harvard University Press.

——— 2012. 'Actions and Activity', *Philosophical Issues*. 22: 233-45.

HUMPHREYS, P. 1989. *The Chances of Explanation: Causal Explanation in the Social, Medical, and Physical Sciences*. Princeton: Princeton University Press.

ISMAEL, J. 2013. 'Causation, Free Will, and Naturalism', in KINCAID, LADYMAN, and ROSS (eds.) *Scientific Metaphysics*. Oxford: Oxford University Press.

——— forthcoming b. 'Against Globalism About Laws', in VAN FRAASSEN, B. and PESCHARD, I. (eds.) *The Experimental Side of Modelling*.

JACKSON, F. and PETTIT, P. 1990. 'Program Explanation: A General Perspective', *Analysis*. 50(2): 107-17.

JONG, H. and SCHOUTEN, M. 2005. 'Ruthless Reductionism: A Review Essay of John Bickle's *Philosophy and Neuroscience: A Ruthlessly Reductive Account*', *Philosophical Psychology*. 18(4): 473-86.

JOYCE, R. 2006. *The Evolution of Morality*. MIT Press.

JUDISCH, N. 2008. 'Why "non mental" won't work: on Hempel's dilemma and the characterisation of the "physical"', *Philosophical Studies*. 140(3): 299-318.

KANE, R. 1989. 'Two Kinds of Incompatibilism', *Philosophy and Phenomenological Research*. 50(2): 219-54.

——— 1996. *The Significance of Free Will*. Oxford: Oxford University Press.

——— 1999. 'Responsibility, Luck, and Chance: Reflections on Free Will and Indeterminism.' *The Journal of Philosophy*. 96(5): 217-40.

——— (ed.) 2002. *Free Will*. Oxford: Blackwell.

——— 2012. 'Torn decisions, luck, and libertarian free will: comments on Balaguer's free will as an open scientific problem', *Philosophical Studies*.

KENDLER, K. and PARNAS, J. (eds.) 2008. *Philosophical Issues in Psychiatry: Explanation, Phenomenology, and Nosology*. Johns Hopkins University Press.

KIM, J. 1992. "'Downward causation" and emergence', in *Emergence or Reduction?: essays on the prospects of nonreductive physicalism*, pp. 119-38. BECKERMANN, A., FLOHR, H., and KIM, J. (eds.), Walter de Gruyter.

——— 1993. *Supervenience and Mind: Selected Philosophical Essays*. Cambridge University Press.

——— 1995. *Supervenience and Mind: Selected Philosophical Essays*. Cambridge: Cambridge University Press.

——— 1995a. 'The nonreductivist's troubles with mental causation', in *Supervenience and Mind: Selected Philosophical Essays*, pp. 336-57. Cambridge: Cambridge University Press.

——— 1995b. "Postscripts on mental causation", in *Supervenience and Mind: Selected Philosophical Essays*, pp. 358-67. Cambridge: Cambridge University Press.

——— 1998. *Mind in a Physical World*. Cambridge, MA: MIT Press.

——— 2002. 'The Layered Model: Metaphysical Considerations', *Philosophical Explorations*. V(I): 2-20.

——— 2003a. 'The American Origins of Philosophical Naturalism', *Journal of Philosophical Research*. Vol. 28, supplement: 83-98.

——— 2003b. 'Blocking Causal Drainage and other Maintenance Chores with Mental Causation', *Philosophy and Phenomenological Research*. LXVII (1): 151-76.

——— 2005. *Physicalism, or Something Near Enough*. Oxford: Princeton University Press.

——— 2010. 'Causation and Mental Causation', in KIM, J. *Essays in the Metaphysics of Mind*, pp. 243-62. Oxford: Oxford University Press.

KINCAID, H. LADYMAN, J. and ROSS, D. (eds.) 2013. *Scientific Metaphysics*. Oxford: Oxford University Press.

KOCH, C. and CRICK, F. 2001. 'The Zombie Within', *Nature*. 411: 893.

KORSGAARD, C. M. 2009. *Self-Constitution: Agency, Identity, and Integrity*. Oxford: Oxford University Press.

LADYMAN, J. 2008. 'Structural Realism and the Relationship between the Special Sciences and Physics', *Philosophy of Science*. 75(5): 744-55.

——— and ROSS, D. 2007. *Every Thing Must Go*. Oxford: Oxford University Press.

LEVY, N. 2005. 'Contrastive Explanations: A Dilemma for Libertarians', *Dialectica*. 59(1): 51-61.

——— 2013. 'Are We Agents at All? Helen Steward's Agency Incompatibilism', *Inquiry*. 56(4): 386-99.

LEWIS, D. 1986. *On the Plurality of Worlds*. Oxford: Blackwell.

——— 1986b. "Causal Explanation." in *Philosophical Papers*. (Vol. 2): 214-40. New York: Oxford University Press.

LIBET, B. 1985. 'Unconscious cerebral initiative and the role of conscious will in voluntary action', *Behavioural and Brain Sciences*. 8(4): 529-39.

LIPTON, P. 1990. 'Contrastive Explanation', in KNOWLES, D. (ed.) *Explanation and Its Limits*. 247-66. Cambridge: Cambridge University Press.

——— 1991. *Inference to the Best Explanation*. London: Routledge.

LIST, C. and MENZIES, P. 2009. 'Nonreductive Physicalism and the Limits of the Exclusion Principle', *Journal of Philosophy*. 106: 475-502.

LOCKE, J. 1689. *An Essay Concerning Human Understanding*. NIDDITCH (ed.) Oxford: Oxford University Press.

LOWE, J. 1996. *Subjects of Experience*. Cambridge: Cambridge University Press.

——— 2002. *A Survey of Metaphysics*. Oxford: Oxford University Press.

——— 2008. *Personal Agency*. Oxford: Oxford University Press.

MARKOSIAN, N. 2012. 'Agent causation as the solution to all the compatibilist's problems', *Philosophical Studies*. 157: 383-98.

——— 1999. 'A compatibilist version of the theory of agent causation', *Pacific Philosophical Quarterly*. 80: 257-77.

MASLEN, C., HORGAN, T., and HABERMANN, H. 2009. 'Mental Causation', in BEEBEE, H., HITCHCOCK, C., and MENZIES, P. (eds.) *The Oxford Handbook of Causation*. Oxford: Oxford University Press.

MELE, A. 1999a. 'Ultimate Responsibility and Dumb Luck', *Social Philosophy and Policy*. 16: 274-93.

——— 1999b. 'Kane, luck, and the significance of free will', *Philosophical Explorations*. 2: 96-104.

MELNYK, A. 1997. 'How to keep the "physical" in physicalism', *The Journal of Philosophy*. 94(12): 622-37.

——— 1999. 'Supercalifragilisticexpialidocious', *Nous*. 33(1): 144-54.

——— 2003. *A Physicalist Manifesto*. Cambridge: Cambridge University Press.

MENZIES, P. 2013. 'Mental Causation in the Physical World', in *Mental Causation and Ontology*. GIBB, S., LOWE, E. J., and INGTHORSSON, R. D. (eds.) 58-87. Oxford: Oxford University Press.

——— and PRICE, H. 1993: 'Causation as a Secondary Quality', *British Journal for the Philosophy of Science*. 44: 187–203.

LA METTRIE, J. O. 1996 [1747]. *La Mettrie: Machine Man and Other Writings*. Ann Thomson (ed. trans.) Cambridge University Press.

MILNER, D. and GOODALE, M. 2004. *Sight Unseen: An Exploration of Conscious and Unconscious Vision*. Oxford: Oxford University Press.

——— 2006. *The Visual Brain in Action*. (Second Edition). Oxford: Oxford University Press.

MOLE, C. 2009. 'Illusions, demonstratives, and the zombie action hypothesis', *Mind*. 118(472): 995-1011.

——— 2013. 'Embodied Demonstratives: A Reply to Wu', *Mind*. 122(485): 231-39.

MONTERO, B. 1999. 'The Body Problem', *Nous*. 33(2): 183-200.

MOORE, G. E. 1922. *Philosophical Studies*. London: Routledge.

MUSALLAM, S., CORNEIL, B. GREGER, B., SCHERBERGER, H., and ANDERSEN, R. 2004. 'Cognitive Control Signals for Neural Prosthetics', *Science* 305: 258–62

NADLER, S. 2008. 'Spinoza and Consciousness', *Mind*. 117: 575-601.

NAGEL, T. 1965. 'Physicalism', *The Philosophical Review*. 74(3): 339-56.

——— 1979. 'Panpsychism', in NAGEL, T. *Mortal Questions*. Cambridge: Cambridge University Press.

——— 1986. *The View from Nowhere*. Oxford: Oxford University Press.

——— 1999. 'Conceiving the Impossible and the Mind-Body Problem', *Philosophy*. 73(285): 337-52.

NAHMIAS, E. 2006. 'Folk Fears about Freedom and Responsibility: Determinism vs. Reductionism', *Journal of Cognition and Culture*. 6(1-2): 215-237.

——— 2010. 'Scientific Challenges to Free Will', O'CONNOR and SANDIS (eds.) *A Companion to the Philosophy of Action*. Wiley-Blackwell.

NAHMIAS, E., MORRIS, S., NADELHOFFER, T. and TURNER, J. 2004. 'The Phenomenology of Free Will', *Journal of Consciousness Studies*. 11(7-8): 162-79.

——— 2005. 'Surveying Freedom: Folk Intuitions about free will and moral responsibility', *Philosophical Psychology*. 18(5): 561-84.

——— 2006. 'Is Incompatibilism Intuitive?', *Philosophy and Phenomenological Research*. 73(1): 28-53.

NAHMIAS, E. COATES, J. and KVARAN, T. 2007. 'Free Will, Moral Responsibility, and Mechanism: Experiments on Folk Intuitions', *Midwest Studies in Philosophy*. 31: 214-42.

NELKIN, D. K. 2011. *Making Sense of Freedom and Responsibility*. Oxford: Oxford University Press.

NICHOLS, S. 2004. 'The Folk Psychology of Free Will: Fits and Starts', *Mind & Language*. 19(5): 473-502.

——— 2006. 'Folk Intuitions about Free Will', *Journal of Cognition and Culture*. 6.

——— 2014. 'Process Debunking and Ethics', *Ethics*. 124(4): 727-49.

NOZICK, R. 1981. *Philosophical Explanations*. Harvard: Harvard University Press.

——— 1993. *The Nature of Rationality*. Princeton University Press.

O'BRIEN, L. and SOTERIOU, M. (eds.) 2009. *Mental Actions*. Oxford: Oxford University Press.

O'CONNOR, T. 1995. 'Agent Causation', in O'CONNOR, T. (ed.) *Agents, Causes, Events*, pp. 173-200. New York: Oxford University Press.

——— 2000. *Persons and Causes: The Metaphysics of Free Will*. Oxford: Oxford University Press.

O'SHAUGHNESSY, B. 1980. *The Will*. (2 vols.) Cambridge: Cambridge University Press.

——— 1992 'The Diversity and Unity of Action and Perception', in CRANE, T. (ed.) *The Contents of Experience*. 216-66. Cambridge: Cambridge University Press.

OLSON, E. 2007. *What Are We? A Study in Personal Ontology*. Oxford University Press.

PARFIT, D. 1984. *Reasons and Persons*. Oxford: Clarendon Press.

PAPINEAU, D. 1990. 'Why Supervenience?', *Analysis*. 50(2): 66-71.

——— 1998. 'Mind the Gap', *Nous*. 32(sup): 373-88.

——— 2002. 'The Rise of Physicalism', in LOEWER and GILLET (eds.) *Physicalism and Its Discontents*. Cambridge: Cambridge University Press.

——— 2008. 'Must a Physicalist be a Microphysicalist?', in HOHWY, J. and KALLESTRUP, J. (eds.) *Being Reduced: New Essays on Reduction and Explanation in the Special Sciences*. Oxford: Oxford University Press.

——— 2009. 'The Causal Closure of the Physical and Naturalism', in MCCLAUGHLIN, et. al. (eds.) *The Oxford Handbook of Philosophy of Mind*. Oxford: Oxford University Press.

——— 2013. 'Causation is macroscopic but not irreducible', in GIBB, S., LOWE, E. J., and INGTHORSSON, R. D. (eds.) *Mental Causation and Ontology*. 126-54. Oxford: Oxford University Press.

——— and SPURRETT, D. 1999. 'A note on the completeness of "physics"', *Analysis*. 59: 25-29.

——— and MONTERO, B. 2005. 'A defence of the *via negativa* argument for physicalism', *Analysis*. 65.3: 233-7.

PEACOCKE, C. A. B. 1979. *Holistic Explanation: Action, Space, Interpretation*. Oxford: Clarendon Press.

PEARL, J. 2000. *Causality*. New York: Cambridge University Press.

PEREBOOM, D. 2001. *Living Without Free Will*. Cambridge University Press.

——— 2014. *Free Will, Agency, and Meaning in Life*. Oxford University Press.

POCKETT, S., BANKS, W. P., and GALLAGHER, S. 2006. *Does Consciousness Cause Behaviour?* The MIT Press.

PRICE, H. 1994. 'Psychology in Perspective', *Philosophy in Mind: Philosophical Studies Series*. 60: 83-98.

PUTNAM, H. 1967. 'Psychological Predicates', in CAPITAN and MERRILL (eds.) *Art, Mind, and Religion*, pp. 37-48. Pittsburgh: University of Pittsburgh Press.

——— 1988. *Representation and Reality*. Cambridge, MA: The MIT Press.

——— 1997. 'Philosophy and our mental life', in *Philosophical Papers Vol. 2*. (291-303) Cambridge: Cambridge University Press.

QUINE, W. V. O. 1948/1949. 'On What There Is', *Review of Metaphysics*. (2): 21-38.

——— 2013. *Word and Object*. (New edition). Cambridge, MA: The MIT Press.

RAATIKAINEN, P. 2010. 'Causation, Exclusion, and the Special Sciences', *Erkenntnis*.

ROSS, D. 2000. 'Rainforest Realism: A Dennettian Theory of Existence', in ROSS, D. BROOK, A. AND THOMPSON, D. 2000. *Dennett's Philosophy: A Comprehensive Assessment*, pp. 147-68. MIT Press.

——— and SPURRETT, D. 2004. 'What to say to a skeptical metaphysician: A defence manual for cognitive and behavioral scientists', *Behavioural and Brain Sciences*. 27: 603-47.

——— and SPURRETT, D. 2007. 'Notions of Cause: Russell's thesis revisited', *British Journal of Philosophy of Science*. 58(1): 45-76.

ROSS, D. BROOK, A. AND THOMPSON, D. 2000. *Dennett's Philosophy: A Comprehensive Assessment*. MIT Press.

SALMON, W. C. 1971. 'Statistical Explanation', in SALMON, W. (ed.) *Statistical Explanation and Statistical Relevance*. 29–87. Pittsburgh: University of Pittsburgh Press.

——— 1984. *Scientific Explanation and the Causal Structure of the World*. Princeton: Princeton University Press.

SANDIS, C. and D'ORO, G. (eds.) 2013. *Reasons and Causes: Causalism and Anti-Causalism in the Philosophy of Action*. Basingstoke: Palgrave Macmillan.

SCHAFFER, J. 2001. 'Physical Causation', *British Journal for the Philosophy of Science*. 52: 809-13.

——— 2004. 'Causes Need Not be Physically Connected to their Effects', in HITCHCOCK, C. (ed.) *Contemporary Debates in Philosophy of Science*, pp. 197-216. Oxford: Blackwell.

——— 2016. 'Cognitive Science and Metaphysics: Partners in Debunking', in MCLAUGHLIN, B. and KORNBLITH, H. (eds.) *Goldman and His Critics*. 337-68. Blackwell.

SCHNALL, S., HAIDT, J., CLORE, G., and JORDAN, A. 2008. 'Disgust as Embodied Moral Judgment', *Personality and Social Psychology Bulletin*. 34(8): 1096-109.

SEAGER, W. 1990. 'Instrumentalism in Psychology', *International Studies in the Philosophy of Science*. 4(2): 191-203.

——— 2010. 'Real Patterns and Surface Metaphysics', in ROSS, D. BROOK, A. AND THOMPSON, D. 2000. *Dennett's Philosophy: A Comprehensive Assessment*, pp. 95-130. MIT Press.

SEARLE, J. 1984. *Minds, Brains and Science*. London: Penguin.

SEHON, S. 2000. 'An Argument Against the Causal Theory of Action Explanation', *Philosophy and Phenomenological Research*. 60(1): 67-85.

SHAPIRO, L. 2012. 'Mental Manipulations and the Problem of Causal Exclusion', *Australasian Journal of Philosophy*. 90(3): 1-18.

SHAPIRO, L. and SOBER, E. 2007. 'Epiphenomenalism. The Dos and Don'ts', in WOLTERS and MACHAMER (eds.) *Thinking About Causes*. 235-64. Pittsburgh: University of Pittsburgh Press.

——— 2012. 'Against Proportionality', *Analysis*. 72(1): 89-93.

SHOEMAKER, S. 2007. *Physical Realisation*. Oxford: Oxford University Press.

SINGER, P. 2005. 'Ethics and Intuitions', *The Journal of Ethics*. 9: 331-52.

——— 2011. *The Expanding Circle: Ethics, Evolution, and Moral Progress*. (Revised ed.) Princeton University Press.

SMART, J.J.C. 1959. 'Sensations and Brain Processes', *Philosophical Review*. 68: 141-56.

SOBER, E. 2001. 'Venetian Sea Levels, British Bread Prices, and the Principle of the Common Cause', *British Journal of Philosophy of Science*. 52: 331-46.

SPIRITES, P., GLYMOUR, C. and SCHEINES, R. 1993. *Causation, Prediction and Search*. New York: Springer-Verlag.

STEWART, H. 1997. *The Ontology of Mind*. Oxford: Clarendon Press.

——— 2008. 'Fresh Starts', *Proceedings of the Aristotelian Society*. 108: 197-217.

——— 2009. 'Sub-intentional Actions and the Over-mentalization of Agency', in SANDIS, C. (ed.) *New Essays on the Explanation of Action*. 295-312. New York: Palgrave Macmillan.

——— 2011. 'Agency, Properties, and Causation', *Frontiers of Philosophy in China*. 6(3): 390-401.

——— 2012. *A Metaphysics for Freedom*. Oxford: Oxford University Press.

——— 2012b. 'The Metaphysical Presuppositions of Moral Responsibility', *Journal of Ethics*. 16(2): 241-71.

——— 2012c. 'Actions as Processes', *Philosophical Perspectives*. 26: 273-88.

——— forthcoming. 'Processes, Continuants, and Individuals', *Mind*.

STICH, S. 1983. *From Folk Psychology to Cognitive Science: The Case Against Belief*. Cambridge, MA: Bradford.

——— 2011. *Collected Papers: Volume 1 Mind and Language*. Oxford: Oxford University Press.

STREVEN, M. 2007. 'Review of Woodward, *Making Things Happen*', *Philosophy and Phenomenological Research*. 74(1): 233-49.

STOLJAR, D. 2006. *Physicalism*. London: Routledge.

——— 2009. 'Physicalism', *The Stanford Encyclopaedia of Philosophy*. (Fall 2009). ZALTA, E.N. (ed.) <http://plato.stanford.edu/archives/fall2009/entries/physicalism/>

STRAWSON, G. 2008. *Real Materialism and Other Essays*. Oxford: OUP.

——— 2010. *Mental Reality* (2nd ed). London: MIT Press.

——— 2011. 'Cognitive Phenomenology: Real Life', *Cognitive Phenomenology*. MONTAGUE, M. and BAYNE, T. (eds.) Oxford: Oxford University Press.

STRAWSON, G. (et al). 2006. *Consciousness and its Place in Nature: does physicalism entail panpsychism?* FREEMAN, A. (ed.) Exeter: Imprint.

STRAWSON, P. F. 1962. 'Freedom and Resentment', *Proceedings of the British Academy*. 48: 187–211.

——— 1992. 'Causation and Explanation', in STRAWSON, P. F. 1992. *Analysis and Metaphysics*. 109–32. Oxford: Oxford University Press.

TANCREDI, L. 2007. 'The neuroscience of 'free will'', *Behavioural Sciences and the Law*. 25: 295–308.

THOMSON, J. J. 1976. 'Killing, Letting Die, and the Trolley Problem', *The Monist*. 59(2): 204–17.

TSE, P. U. 2013. *The Neural Basis of Free Will: Criterial Causation*. MIT Press.

TOMBERLIN, J. (ed.) 1997. *Philosophical Perspectives 11: Mind, Causation, and World*. Boston: Blackwell.

TOOLEY, M. 1997. *Time, Tense, and Causation*. Oxford: Clarendon Press.

TYE, M. 2011. *Consciousness Revisited: Materialism Without Phenomenal Concepts*. London: MIT Press.

UNGERLEIDER, L. and MISHKIN, M. 1982. 'Two cortical visual systems', in INGLE, D. J., GOODALE, M., MANSFIELD, R. J. W. (eds.) *Analysis of Visual Behavior*. 549–86. Cambridge, MA: The MIT Press.

VAN INWAGEN, P. 1983. *An Essay on Free Will*. Oxford: Oxford University Press.

——— 1998. 'The Mystery of Metaphysical Freedom', in KANE (ed.) *Free Will*. 189-96. Blackwell.

VARGAS, M. 2009. 'Revisionism About Free Will: A Statement and Defence', *Philosophical Studies*. 144(1): 45-62.

——— 2013. *Building Better Beings*. Oxford University Press.

VELLEMAN, J. D. 2000. *The Possibility of Practical Reason*. Oxford: Oxford University Press.

——— 2000. 'What Happens When Someone Acts?', in VELLEMAN, J. D. *The Possibility of Practical Reason*, pp. 123-43.

VIGER, C. 2000. 'Where Do Dennett's Stances Stand? Explaining Our Kind of Mind', in ROSS, D. BROOK, A. AND THOMPSON, D. 2000. *Dennett's Philosophy: A Comprehensive Assessment*, pp. 131-46. MIT Press.

VIHVELIN, K. 2000. 'Libertarian Compatibilism', *Philosophical Perspectives*. 14: 139-66.

——— 2004. 'Free Will Demystified: A Dispositional Account', *Philosophical Topics*. 32(1): 427-50.

——— 2013. *Causes, Laws, and Free Will: Why Determinism Doesn't Matter*. Oxford: Oxford University Press.

VON WRIGHT, G. 2007. *Explanation and Understanding*. New York: Cornell University Press.

WALLHAGEN, M. 2007. 'Consciousness and Action: Does Cognitive Science Support (Mild) Epiphenomenalism?', *British Journal for the Philosophy of Science*. 58: 539-61.

WEGNER, D. 2002. *The Illusion of Conscious Will*. MIT Press.

WIGGINS, D. 1973. 'Towards a reasonable libertarianism', in HONDERICH, T. (ed.) *Essays on Freedom of Action*. London: Routledge & Kegan Paul.

WOODWARD, J. 1984. 'A theory of singular causal explanation', *Erkenntnis*. 21(3): 231-62.

——— 2003. *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.

——— 2007. 'Interventionist Theories of Causation in Psychological Perspective', in GOPNIK, A. and SCHULZ, J. (eds.) *Causal Learning: Psychology, Philosophy and Computation*. 19-36. Oxford: Oxford University Press.

——— 2008. 'Mental causation and neural mechanisms', in HOHWY, J. and KALLESTRUP, J. (eds.) *Being Reduced: New Essays on Reductive Explanation and Special Science Causation*. 218-62. Oxford: Oxford University Press.

——— 2008. 'Cause and Explanation in Psychiatry: An Interventionist Perspective', in KENDLER, K. and PARNAS, J. (eds.) *Philosophical Issues in Psychiatry: Explanation, Phenomenology, and Nosology*. 132-83.

——— 2011. 'Interventionism and Causal Exclusion' [preprint]. Available at: <http://philsci-archive.pitt.edu/8651/>

——— 2011. 'Mechanisms Revisited', *Synthese*. 183(3): 409-27.

——— 2012. 'Causation: Interactions between Philosophical Theories and Psychological Research', *Philosophy of Science*. 79(5): 961-72.

——— 2013. 'Causation and Manipulability', *The Stanford Encyclopedia of Philosophy* (Winter 2013 Edition). ZALTA, E. N. (ed.) [<http://plato.stanford.edu/archives/win2013/entries/causation-mani/>]

——— and HORGAN, T. 1985. 'Folk Psychology is Here to Stay', *The Philosophical Review*. XCIV (2): 197-226.

WOOLHOUSE, R. S. 1985/6. 'Leibniz's Reaction to Cartesian Interaction', *Proceedings of the Aristotelian Society*. 86: 69-82.

WORLEY, S. 2006. 'Physicalism and the Via Negativa', *Philosophical Studies*. 131(1): 101-126.

WU, W. 2013. 'The Case for Zombie Agency', *Mind*. 122(485): 217-30.

YABLO, S. 1992. 'Mental Causation', *Philosophical Review*. 101(2): 245-80.

YANG, E. 2012. 'Eliminativism, interventionism, and the Overdetermination Argument', *Philosophical Studies*, pp. 1–20. 10.1007/s11098-012-9856-0.